*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2746–2753
Marseille, 11–16 May 2020


# Exploring a Choctaw Language Corpus with Word Vectors and Minimum Distance Length

**Jacqueline Brixey**[1 3], **David Sides**[2], **Timothy Vizthum**, **David Traum**[3], **Khalil Iskarous**[1]
[1]University of Southern California, [2]Mississippi State University, [3]USC Institute for Creative Technologies
brixey@usc.edu, dhs131@msstate.edu, t.vizthum@gmail.com, traum@ict.usc.edu, kiskarou@usc.edu

**Abstract**

This work introduces additions to the corpus ChoCo (*Choctaw language Corpus*), a multimodal corpus for the American indigenous language Choctaw. Using texts from the corpus, we develop new computational resources by using two off-the-shelf tools: word2vec and Linguistica. Our results indicate these tools need expert input to reliably interpret the results.

**Keywords:** endangered languages, indigenous language, low resource languages, Choctaw, Native American languages

## 1. Introduction

Developing computational resources for a language is a time and labor-intensive undertaking. Many languages in the world are becoming endangered and extinct before resources can be developed. However, several off-the-shelf tools exist that can be used to explore low resource languages that require no expert knowledge nor substantial labor.

In this work, we use two such tools to create new computational resources for Choctaw, an American indigenous language. The first tool is word2vec, which provides lexical understanding of a language without the use of a dictionary or word list. The second tool, Linguistica, utilizes unsupervised learning to detect morphological information.

Our experiments demonstrate that off-the-shelf tools could be used with minimal labor to increase available linguistic resources for low-resource languages. However, expert knowledge was still needed to interpret the results and remove numerous false positives. We suggest how to tune two tools to return more meaningful results for similar languages, thus creating a road map for increasing resources for other under-resourced indigenous languages.

Our other contribution in this paper is to introduce new additions to our previous corpus work for Choctaw (Brixey et al., 2018). We added over 90,000 new tokens of text from diverse sources, as well as an additional dictionary. The corpus contains a diversity of text, audio, and video.

## 2. Choctaw People and Language

The Choctaw language is spoken by the Choctaw tribe, an American indigenous group that originally lived in what today is Alabama and Mississippi. In the early 1830s the Choctaws were forcibly relocated to Oklahoma in the migration known as the Trail of Tears, although some people remained.

Choctaws are the third most populous tribe by tribal group population in the United States, with around 195,000 people identifying as Choctaw in the 2010 US census.[1] There are three federally recognized Choctaw tribes[2]: Jena Band of Choctaw Indians (in Louisiana), Mississippi Band of Choctaw Indians, and the Choctaw Nation of Oklahoma. The first and third authors are both enrolled members of the Choctaw Nation of Oklahoma.

The Choctaw language has a "Threatened" classification according to Ethnologue (Simons and Fennig, 2018), as there are only approximately 10,000 fluent speakers (less than 1% of the Choctaw population), and the language is losing users. For many speakers in Oklahoma (Williams, 1999), Choctaw is their second language, and revitalization efforts have established language courses at local schools and colleges. Choctaw is spoken by all ages in Mississippi (Simons and Fennig, 2018), but is losing speakers over time.

The Choctaw language belongs to the Western Muskogean language family; it is most closely related to Chickasaw (Haas, 1979). Public policies enacted during the 1900s were designed to forcefully assimilate Native Americans and suppress indigenous languages (Battiste et al., 2000). As a result, many Native Americans did not learn their ancestral language, and few works are published in these languages in print or online.

### 2.1. Language Variation and Orthography

The relevant literature has debated if the variation in Choctaw sufficiently indicates different dialects (Broadwell, 2005; Broadwell, 2006; Nicklas, 1972). Sources agree that there are three dialect variants in Mississippi, however it is unclear how those Mississippi variants are represented in Oklahoma, and if there are new variants due to mixing of historical variants. The literature concludes that variation in spoken Choctaw is fairly minor, with some variation in phonetic detail (Ulrich, 1993). Anecdotally, Choctaw speakers in many locations refer to "dialects", however this is typically used to indicate any difference in word choice between speakers. When asked, speakers cannot identify specific dialects, beyond grouping Oklahoma

[1] https://www.census.gov/population/www/cen2010/cph-t/t-6tables/TABLE%20(1).pdf

[2] https://www.govinfo.gov/content/pkg/FR-2017-01-17/pdf/2017-00912.pdf
For a definition of "federal recognition"for US tribes, see https://www.bia.gov/frequently-asked-questions

| Characters (IPA in brackets when different from spelling) | | | |
|---|---|---|---|
| p b t k f [ɸ] s h m n l w y [j] | | | |
| IPA | Traditional | Mississippi | |
| | | School | Modern |
| tʃ | ch | č | ch |
| ʃ | sh | š | sh |
| ɬ | hl, lh | ł | lh |
| a | ʊ, v, a | a | a |
| aː | a | á | á |
| ã | a̲, an, am | ą | a̲ |
| i | i | i | i |
| iː | e, i | í | í, e |
| ĩ | i̲, in, im | į | i̲ |
| o | u, o | o | o |
| oː | o | ó | ó |
| õ | o̲, u̲, on, om | ǫ | o̲ |

Figure 1: Choctaw sounds and orthographic variants.

| English | Oklahoma | Mississippi |
|---|---|---|
| **January** | Koichush Hʊshi | Hoponi Hashi |
| **February** | Watonlak Hʊshi | Hohchafo Iskitini |
| **March** | Mahli Hʊshi | Hohchafo Chito |
| **April** | Tek Ihʊshi | Kowichoo̲sh Hashi |
| **May** | Bihi Hʊshi | Kowichito Hashi |
| **June** | Bissa Hʊshi | Mahlih Hashi |
| **July** | Kafi Hʊshi | Watoo̲lak Hashi |
| **August** | Takkon Hʊshi | Tik I̲hashi |
| **September** | Hoponi Hʊshi | Bihi Hashi |
| **October** | Chafo Iskitini Hʊshi | Bissa Hashi |
| **November** | Hohchafo Chito Hʊshi | Hashi Kafi |
| **December** | Koichito Hʊshi | Takkon Hashi |

Table 1: Months in Oklahoma and Mississippi variants

speakers separate from Mississippi speakers.

In this paper, we will refer to the different language groups as "variants". Broadwell (2006) identifies four present-day regional variants: Mississippi Choctaw, Oklahoma Choctaw, Louisiana Choctaw, and Mississippi Choctaw of Oklahoma (spoken by Choctaws who live in the Chickasaw Nation in Oklahoma).

There is a vast amount of variation between the groups concerning orthographic conventions. Three prominent variants (corresponding to some of the systems described by Broadwell (2006, section 1.2)) are shown in Figure 1. The Oklahoma variant today uses the traditional orthography, while the Mississippi orthography has used several writing systems.

An example of the difference in both orthography and lexical items is found in the naming of months, shown in Table 1.

### 2.2. Overview of the Choctaw language

Choctaw word order is subject-object-verb, and adjectives follow the nouns they modify. Choctaw has complex morphology, including infixes, prefixes, and suffixes that can occur in combinations together. Allomorphy and vowel reduplication also occur.

The following examples (in the Oklahoma orthography) illustrate some of the morphology features of Choctaw; the second line in each example shows a morpheme in angle brackets. The first example shows h-grade infixing and vowel reduplication in the verb *tahakchi* ("to tie quickly"), which is the inflected form of the base verb *takchi*("to tie"). The second example illustrates that the second singular negation subject pronoun *chik*, and the negation suffix *o*; the affirmative version of this sentence would be *Ish impa.* ("You eat."), as shown in the third example.

1. Ashekonopa ilʊppʊt tahakchi li.
   Ashekonopa ilʊppʊt ta<ha>kchi li.
   knot this tie<quickly> 1SG
   I tie this knot quickly.

2. Chik impo.
   <Chik> imp-<o>.
   not.2SG eat-NEG
   You are not eating.

3. Ish impa.
   <Ish> impa.
   2SG eat
   You are eating.

In the traditional orthography, some morphemes attach to the verb base, while other do not. However, this standard has changed over time. For the Mississippi orthography, the standard was previously to attach all morphemes to the base, however this has also changed over time with more morphemes not agglutinating.

## 3. Data Set

This paper adds new materials to our previous work to create a multimodal Choctaw language corpus (Brixey et al., 2018); the corpus is named ChoCo (*Choctaw language Corpus*, previously named *Chahta Anumpa*). Our additions add over 90,000 new tokens to the corpus.

### 3.1. New Additions

The additions to the corpus include more texts in the Mississippi variant, religious texts, and a recently published dictionary in the Oklahoma orthography (The Choctaw Nation of Oklahoma Dictionary Committee, 2016). We added a stop word list for both variants, as well as sets of prefixes and suffixes for both variants. The stop word lists and affix sets are discussed in Section 4.

#### 3.1.1. Mississippi variant

The files were acquired from a member of the Mississippi Band of Choctaw Indians (MBCI) in Philadelphia, MS. The data consisted of five different printed materials: a Mississippi Choctaw calendar from 1984, the third installment of an anthology produced by the MBCI, a publication entitled

*Choctaw Material Culture*, a children's coloring book entitled "Na Yo Pisa", and a short publication on different jobs in the Mississippi Choctaw community, entitled "Okla Apilachi". All texts are written in the school variant of Mississippi Choctaw, comprising the first addition to the corpus in this orthography. The majority of the files were published through the Choctaw Heritage Press in 1982, with the exception of the calendar. The texts provide a number of insights, for example, a full list of the months in Mississippi Choctaw. When compared with the months from the Oklahoma community, differences in both orthography and month order can be seen. From these files, more than 1,100 words were added to the corpus. Of particular interest are Choctaw translations of the songs "Amazing Grace", "Oh, How I Love Jesus", and "Sweet By and By".

#### 3.1.2. Oklahoma variant

The first addition is a bilingual Choctaw-English dictionary published by the Choctaw Nation of Oklahoma (The Choctaw Nation of Oklahoma Dictionary Committee, 2016). At present, we added only the Choctaw-English portion of the dictionary. The dictionary was available for free online as a pdf [3]. The dictionary file was first run through a generic free online OCR [4], then converted into an Excel file from the doc file produced by the OCR website. Next, we manually corrected errors produced during the OCR process. Common errors found in the OCR correction process included the replacement of the letter "h" with the letters "fl", the replacement of the letter "u" with the letter "v", and the omission or misplacement of nasal diacritic. The dictionary contains over 5000 entries.

Additional phrases from the Choctaw Nation of Oklahoma's newspaper Biskinik were added. The corpus previously contained 539 Choctaw tokens from the newspaper, it now contains 4867 tokens. We downloaded all pdf editions available from the newspaper's archives online [5] and then manually extracted any phrases written in Choctaw. Nearly all of the phrases came from language lessons in the newspaper, few phrases came from in-line text. All of the text from Biskinik are in Choctaw and in English. The extracted phrases were saved in an Excel file. The corpus contains the Excel file and a plain text version of the Excel file.

There are five columns in the Excel file: a column lists the originating newspaper file name, a column cites the page where the phrases were found, the phrases in Choctaw and English are in two respective columns, and a column for notes. The notes column lists any relevant details. For example in one lesson, the texts were given in Choctaw, and the English translation was given in a subsequent edition. The corpus contains both the Excel file and a plain text tab-delimited version of the Excel file.

We also added two religious texts in the traditional orthography, The Book of Psalms [6] and The Book of Joshua [7]. Both texts were freely available online. Both had been previously converted into text, however contained a number of errors as a result of generic OCR processes. We manually corrected these errors. The Book of Psalms contains a total of 58,648 word tokens, while The Book of Joshua contains 21,948 tokens.

| Type | Oklahoma | Mississippi |
|---|---|---|
| Short stories | 5954 | 1693 |
| Phrases | 17039 | 1431 |
| Poetry | 243 | — |
| Correspondence | 159 | — |
| Religious texts | 80818 | 30010 |
| Examples from scholarly sources | 589 | 12 |
| Monolingual texts | 1344 | 41 |
| Dictionaries | 46704 | — |
| Total | 152850 | 32087 |

Table 2: Word token counts in Choctaw for texts in two variants

### 3.2. Content and Format

ChoCo contains audio, text, and video in the Choctaw language.

All text is stored in a separate folder within the data set. A brief description of the available types of text are shown in Table 2. Within the text folder in the data set, there are seven Excel files for monolingual texts, poetry, short stories for both variants, short phrases for both variants, and correspondence. Within the text folder, there are four subfolders. The first subfolder is for dictionaries. All dictionaries are stored as separate Excel files. Biskinik, the newspaper, also has a subfolder, with a subfolder for all original pdfs. Text pulled from Biskinik phrases are stored as an Excel file and plain text version. The third subfolder contains religious texts that were manually corrected, each text is a .doc and text file. The final subfolder contains all the scanned images for the Mississippi variant (described in detail in Section 3.1.1).

The audio folder of the corpus contains 12 files of spoken and sang Choctaw from the Global Recordings Network[8] in one subfolder. There is a second subfolder which contains all the audio files scraped from the Lesson of the Day from the School of Choctaw Language website[9]. Each audio clip is approximately thirty seconds long. All clips were transcribed by the School of Choctaw Language, the transcriptions are in the text subfolder.

[3] http://www.choctawschool.com/media/369055/New%20Choctaw%20Dictionary.pdf

[4] https://pdf2docx.com/

[5] https://www.choctawnation.com/biskinik-newspaper-archive

[6] https://archive.org/details/bookofpsalmstran00wrig/page/n4

[7] https://archive.org/details/booksofjoshuajud00wrig/page/n8

[8] http://globalrecordings.net/en/program/4680

[9] http://www.choctawschool.com/lesson-of-the-day.aspx

Finally, the video folder contains 30 mp4 video files downloaded from YouTube that contain Choctaw speech (Brixey et al., 2018). The videos were manually annotated with a category type: Songs (12 videos), Cultural (3 videos), Story (1 video), and Instructional (14 videos). No videos have been transcribed, we leave this to future work.

### 3.3. Availability

The corpus in its current form has been shared with the Mississippi Band of Choctaw Indians (MBCI) to support documentation and revitalization efforts. As more texts are added to the corpus, new versions of the corpus will be submitted to the MBCI for archiving.

Following the OCR correction of all available religious texts, we will submit the corpus to the Sam Noble Museum archives[10] for permanent storage and where it can be obtained by researchers upon request.

## 4. Exploration of the data set

We explored the text data set using Linguistica and with word2vec. All experiments were conducted on Oklahoma texts and Mississippi texts separately, we report the results for each variant.

### 4.1. Linguistica

Linguistica is a tool that can serve as a first step to creating affix lists for a language (Lee and Goldsmith, 2016). It successfully detects morphemes for European languages, such as French, that do not have a high average number of morphemes per word, and reportedly works best for words under 6-grams (Goldsmith, 2001, p.172).

The input to Linguistica is one text file, in this case, one text file containing all texts for the Mississippi variant, and one text file for the Oklahoma variant. Linguistica's unsupervised learning of morphology is achieved using Minimum Description Length (MDL), an algorithm that finds the best description of a set of data by finding the model that compresses the data best. MDL in Linguistica works to find the minimum number of morphological patterns needed to describe the given corpus (Goldsmith, 2006).

One key step in the algorithm is to discover morpheme boundaries in order to create affix and stem sets. To find the split, Linguistica uses two heuristics in which the first feeds into the second (Goldsmith, 2001). The first is successor frequency, which works by calculating how frequently a given character follows another character in a string. Peaks in successor frequency indicate morpheme splitting sites. One example is if given the string *gover* in English, the successor frequency of the letter *n* following this string is 1, no other letters could possibly follow. However, if we have the string *govern*, then the success frequency is 6, as *governed, governing, government, governor, governs*, and *govern* could be acceptable inflected forms. The second heuristic is based on the probability that a suffix length might only be one letter, as suffixes of only one letter are rare and unusual. The tendency then leans towards suffixes being at least two letters long (Goldsmith, 2006).

The results of Linguistica for Choctaw would be an indicator that the tool can be used on morphologically-rich languages. Many indigenous languages have little to no resources, but successful results from Linguistica for Choctaw would indicate that this tool could be used to easily generate affix list resources for other morphologically-rich languages.

We hypothesised that Linguistica would pick up on suffixes of greater than three letters for both variants, such as *-chi* (a causative suffix). It is a frequently occurring suffix. We did not expect it to pick up on *-o* (negation suffix) as this is a one letter suffix; the documentation on the language indicated the algorithm is steered away from one letter suffixes. The Mississippi orthography tends to stack more morphemes on the stem than the Oklahoma orthography does, thus the Mississippi variant results should contain more affixes than the Oklahoma. This is because the Oklahoma variant affixes will look more like standalone words, rather than affixes. Morphemes do not behave like words, however, and must follow a strict placement order around the verb or noun base (Broadwell, 2017).

#### 4.1.1. Linguistica methods

Linguistica was run using default parameters. The data were not preprocessed. The newest version of Linguistica, Linguistica 5.2.1 [11] only detects suffixes, thus we used Linguistica 3[12], as it also discovers prefixes.

#### 4.1.2. Linguistica results

The full results are in Table 3. For the Mississippi text, Linguistica found 144 prefixes and 27 suffixes. For both prefixes and suffixes, the most common affix Linguistica discovered was the null affix. For prefixes, only five of the 144 proposed prefixes were true prefixes. In total, it detected thirteen correct suffixes. The three most common suffixes it found was *-h* (a suffix indicating the present tense), *-t* (a suffix used for verb phrases), and *tók* (a suffix indicating the past tense). It was unusual that -h and -t were detected since these are one letter morphemes, since the Linguistica algorithm is designed to detect morphemes with a length of greater than two letters.

For the Oklahoma variant, Linguistica returned 95 prefixes and twelve suffixes. Of the 95 proposed prefixes, only two were real affixes in the language. Most of the returned prefixes are simply the first three letters in various words that do not carry any meaning. Six of the twelve detected suffixes are correct.

Linguistica did show the ability to detect some correct prefixes and suffixes. However, the high false-positive rate (for example, two correct prefixes out of a total of 95 suggested prefixes) required expert knowledge to eliminate incorrect affixes from the returned lists. As described previously, the Mississippi orthography tends to show morphemes attached to the stem, while the Oklahoma variant tends to space morphemes. For this reason, the Mississippi results returned

[10] https://samnoblemuseum.ou.edu/collections-and-research/native-american-languages/

[11] http://linguistica-uchicago.github.io/lxa5/index.html

[12] http://people.cs.uchicago.edu/~jagoldsm/linguistica-site/downloads.html

| | Oklahoma | Mississippi |
|---|---|---|
| Prefixes | sʊ | a̲ |
| | pi̲ | im |
| | | chi |
| | | chim |
| | | hapi |
| Suffixes | h | h |
| | chi | t |
| | o | ósh |
| | shke | chi̲ |
| | t | at |
| | ma̲ | ka̲ |
| | | chi |
| | | ttók |
| | | yat |
| | | ho̲ |
| | | nnah |
| | | li |
| | | tók |

Table 3: Prefixes and suffixes detected by Linguistica for OK and MS variants, excluding the NULL affix

more affixes. Overall, the approach of Linguistica is better for detecting suffixes than prefixes for both variants.

## 4.2. word2vec

Creating word vectors is often a preprocessing step towards using a text data set for many tasks, such as giving the word vectors to a machine learning model to make predictions. However, the process of creating the word vectors maps semantically related words. By exploring the mappings, we can potentially discover relationships between words and other meaningful syntactic and semantic information captured by the word vectors (Mikolov et al., 2013b). This can be highly useful when exploring a language with little documentation as meanings from novel words can be derived from similar known words.

### 4.2.1. word2vec methods

We constructed the word vectors using Gensim word2Vec (Řehůřek and Sojka, 2010), an open-source tool that processes text to create a neural network. The resulting neural network is a set of feature vectors of words in the corpus represented numerically, and a set of vectors containing the probabilities that those words will co-occur. Vectors of similar words group together in the vector space, so that it is possible to predict a given word's meaning based on the neighboring vectors (Mikolov et al., 2013a).

We experimented with word2vec in three ways. First, we tuned only the parameters. Fine-tuning hyperparameters of a neural network is a means to improve performance. Second, we removed stop words - words that occur in high frequencies but do not add additional meaning. Stop words are often determiners, articles, and subject pronouns. The list of stop words is a text document also contained in the corpus. Third, we wrote a stemmer script that removes inflections from stems based on the affix sets discovered by Linguistica (discussed in Section 4.1.).

We evaluated the resulting word embeddings for all experiments through two methods. First, we made word plot images. We visually inspected these images for relationships within word clusters. We also conducted most-similar-words and analogy word tests.

The analogy task consists of two pairs of words that share a relation, and the last word of the second pair is inferred based on the other three. The relation between words in a pair is not explicitly given, the relation must first be determined and then applied to the second pair (Levy and Goldberg, 2014). The most common example given in this task is "man is to woman, as king is to...?", with the correct answer being "queen".

Nearly all of our word tests examine the encyclopedic knowledge of the word vectors (Gladkova et al., 2016). We evaluated how relevant the results were for each word test. The tests include four tests to find similar words, and three analogy tests. To design the analogy tests, we referred to a taxonomy of semantic relations (Bejar et al., 2012). The word tests are:

1. Three most common words to *holisso* (book)
2. Three most common words to *tek* (female)
3. Three most common words to *tuklo* (two)
4. Three most common words to *tohbi* (white)
5. The analogy "woman is to mother as man is to ...?"
6. The analogy "fat is to skinny as old is to ...?"
7. The analogy "hunger is to eat as thirst is to ...?"

### 4.2.2. word2vec Oklahoma results

The graphs of the mapped vectors, Figures 2-5 show that removing the stop words does reduce the amount of noise. Dimensionality reduction to produce all graphs was achieved using PCA (Principal Components Analysis). For example in Figures 2 and 4, many subject pronouns are visible (such as *is, ish, sʊ, si, sa, in, li*). Once stop words are removed, we can see differentiation between nouns and verbs in Figures 3 and 5.

For the most-similar-word tests, the most successful word test was number 3, to find words similar to the number "two". The configuration that returned the best results, in this case eight correct words out of ten possible, contained both stop words and inflections, and used the defaults settings for the skipgram model.

The first most-similar-word test found the words "desk" and "paper" using the skipgram model. Nearly all of the models found the word "stone" as a similar word to "book", often returning it as the most similar word. We suspect that this might be because many texts composing the data set are religious in nature, and described text as having been written on stone.

For test two, we examined if the model could return "female" as an adjective or as a noun, as both are acceptable uses of the word in Choctaw. Overall, the models returned more nouns, such as "wife" (stemmed text with default parameters), than adjectives. One unexpected result was that the system returned female names, such as "Sue" and "Joyce", for this test.

It is unclear why the fourth test, to return other color words, was unsuccessful for all of the models. It is also unclear

Table 4: Plots for OK data

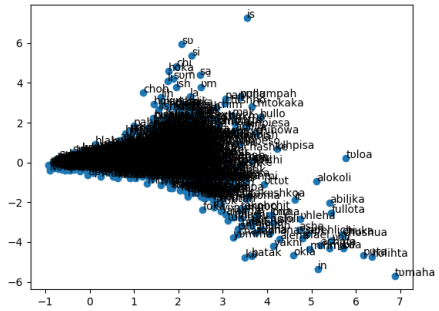

Figure 2: Plots for data with no stop words or stems removed, OK text

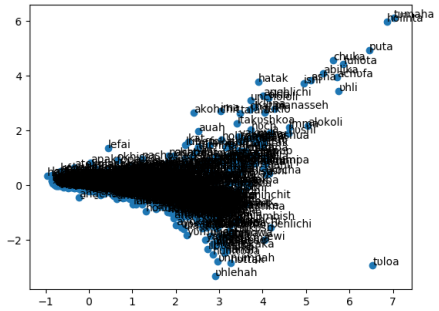

Figure 3: Plots for data with stop words removed, OK text

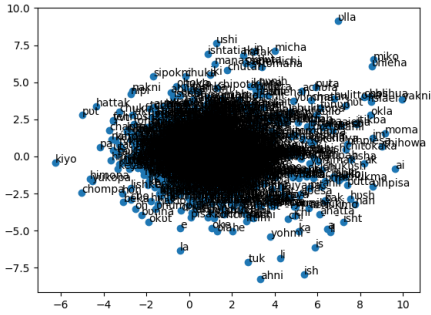

Figure 4: Plots for data with stemmer applied, OK text

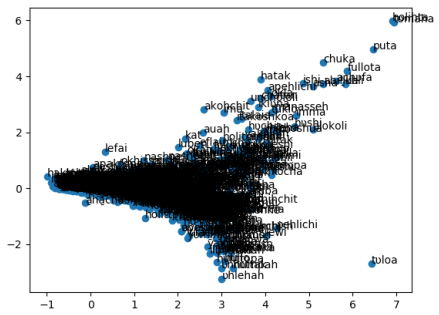

Figure 5: Plots for data with stemmer applied, OK text

why the first analogy test, which should have returned "father", was only able to return the closest words meaning "head of" (run with vector dimension = 300 and epochs = 20) and "captain" (run with skipgram model).

The skipgram model returned acceptable results for the second analogy as well. It found "young" when given stemmed text. The word for "youth" was also found (epochs= 20, vector dimension= 300.

The final analogy test also led to mixed results, with no correct response "to drink" produced. The word "to cook" was one produced answer, as was a second verb that means "to eat". The closest word found by the vectors to the act of drinking was the word for "throat". However, this word is a noun rather than a verb.

#### 4.2.3. word2vec Mississippi results

Figures 6 through 9 show the mappings for the vectors trained on the Mississippi variant text data. We can see that the shapes are similar to those of the Oklahoma results, and that there is similar noise in the figures that include stop words (Figure 6 and 8). Figure 7 is similar to Figure 4 in that applying the stemmer alters the plotting of the vectors more than any other parameter change.

Since the Mississippi text file contained fewer tokens, many of the word tests could not be reused because the words were not present. Word tests 1, 3, 4, and 5 were evaluated for the Mississippi data.

1. Three most common words to *holisso* (book)
3. Three most common words to *tuklo* (two)
4. Three most common words to *tohbi* (white)
5. The analogy "woman is to mother as man is to ...?"

One interesting result from word test 1 is that with the default parameters and no alteration to the text, the second most common word was "Jesus". We suspect this is because the data are largely composed of religious text.

For the word tests, the most successful word test was again number 3, to find numbers. The configuration that returned the best results, which was two numbers out of a possible ten results, applied the stemmer to the text, had a vector dimension of 300, and twenty epochs. The least successful configurations in this case were vectors that used the data with stop words removed.

## 5. Discussion

The goal of these experiments was to find settings for off-the-shelf tools that can be used on low-resource languages to increase available linguistic resources for those languages with minimal labor and expert knowledge. Using text data for the language Choctaw, we experimented with the tools Linguistica and word2vec.

For Linguistica, we found that we could discover limited sets of prefixes and suffixes for the language with the default parameters of the tool. While it has no settings to discover infixes in the language, Linguistica nevertheless produced meaningful lists for both prefixes and suffixes for both variants of the language. We used these lists to create a stemming script, which then improved the performance of

Table 5: Plots for OK data

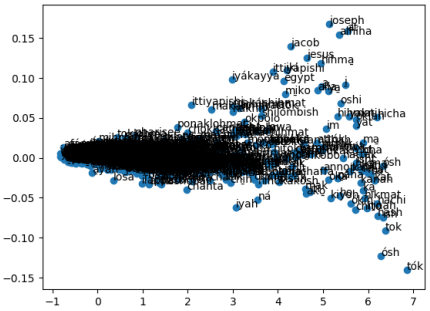

Figure 6: Plots for data with no stop words or stems removed, MS text

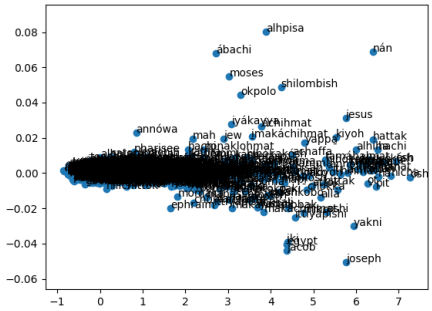

Figure 7: Plots for data with stop words removed, MS text

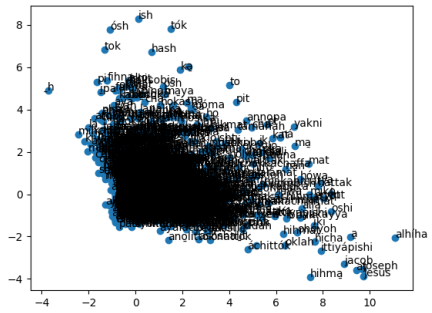

Figure 8: Plots for data with stemmer applied, MS text

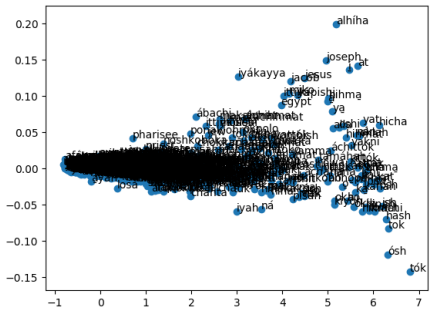

Figure 9: Plots for data with stop words removed and stemmed, MS text

word2vec. However, the results contained numerous errors, and it was necessary to have expert knowledge to determine if a suggested affix was a false-positive. It is possible this tool could be successfully used by non-experts on other languages if the data set is larger and/or the given language's morphology is less complex.

The tool word2vec creates lexical representation for text in the data set. We experimented with altering the parameters, removing stop words, and stemming inflected words to determine what produced better results in terms of most-similar-words and analogy tests. We determined that the skipgram model was the best single parameter to alter, as it improved on the baseline model for nearly all of the tests. We also determined that increasing the dimensions of the vectors and increasing the number of training epochs from the default parameter settings improved results.

## 6. Conclusion

This paper introduced additions to the multimodal Choctaw language corpus ChoCo. We explored the data using two off-the-shelf resources to gain insights into the lexical and morphological aspects of the language. We found that the two resources, Linguistica and word2vec, did produce meaningful results despite the smallness of the data set and the complexity of the language's morphology. However, expert input was still required. We leave it to future work to determine if a larger data set improves results.

## 7. Bibliographical References


Battiste, M., Youngblood, J., et al. (2000). *Protecting Indigenous knowledge and heritage: A global challenge*. UBC Press.

Bejar, I. I., Chaffin, R., and Embretson, S., (2012). *Cognitive and psychometric analysis of analogical problem solving*, chapter 3, pages 58–64. Springer Science & Business Media.

Brixey, J., Pincus, E., and Artstein, R. (2018). Chahta anumpa: A multimodal corpus of the Choctaw language. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Broadwell, G. A. (2005). Choctaw. In Janine Scancarelli et al., editors, *Native Languages of the Southeastern United States*. U of Nebraska Press.

Broadwell, G. A. (2006). *A Choctaw Reference Grammar*. U of Nebraska Press.

Broadwell, G. A. (2017). Parallel affix blocks in choctaw. In Stefan Muller, editor, *Proceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar*, pages 103–119, University of Kentucky,Lexington. CSLI Publications.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California, June. Association for Computational Linguistics.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Nat. Lang. Eng.*, 12(4):353–371, December.

Haas, M. R. (1979). Southeastern languages. In Lyle Campbell et al., editors, *The Languages of Native America: Historical and Comparative Assessment*, pages 299–326. University of Texas Press.

Lee, J. and Goldsmith, J. (2016). Linguistica 5: Unsupervised learning of linguistic structure. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 22–26, San Diego, California, June. Association for Computational Linguistics.

Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Nicklas, T. D. (1972). *The Elements of Choctaw*. Ph.d. dissertation, University of Michigan.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Gary F. Simons et al., editors. (2018). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-first edition.

The Choctaw Nation of Oklahoma Dictionary Committee. (2016). *Chahta Anumpa Tosholi Himona: New Choctaw Dictionary*. Choctaw Print Services, 1st edition.

Ulrich, C. H. (1993). The glottal stop in western muskogean. *International journal of American linguistics*, 59(4):430–441.

Williams, R. S. (1999). Referential Tracking in Oklahoma Choctaw: Language Obsolescence and Attrition. *Anthropological linguistics*, pages 54–74.