4
Language Documentation & Conservation Special Publication No. 20
Collaborative Approaches to the Challenge of Language Documentation and Conservation: Selected papers from the 2018 Symposium on American Indian Languages (SAIL)
ed. by Wilson de Lima Silva and Katherine Riestenberg, p.38-53
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24931

"Data is nice:" Theoretical and pedagogical implications of an Eastern Cherokee corpus

Benjamin Frey
University of North Carolina, Chapel Hill

This paper serves as a proof of concept for the usefulness of corpus creation in Cherokee language revitalization. It details the initial collection of a digital corpus of Cherokee/English texts and enumerates how corpus material can augment contemporary language revitalization efforts rather than simply preserving language for future analysis. By collecting and analyzing corpus material, we can quickly create new classroom materials and media products, and answer deeper theoretical linguistic questions. With a large enough corpus, we can even implement machine translation systems to facilitate the production of new texts. Although the vast majority of print material in Cherokee is in the Western dialect, this corpus has focused on Eastern texts. Expanding the dataset to include both dialects, however, will allow for comparison and facilitate generalizations about the Cherokee language as a whole. A corpus of Cherokee data can answer second language learners' questions about the structure of the language and provide patterns for more effective, targeted learning of Cherokee. It can also provide teachers with ready access to accurate representations of the language produced by native speakers. By combining documentation and technology, we can leverage the power of databases to expedite and facilitate language revitalization.

1. INTRODUCTION. The use of corpus material for language teaching and linguistic analysis is not new. Reppen (2010) cites many examples of the usefulness of corpora in the language classroom, while I and several colleagues have shown the potentially paradigm-shifting value of a thorough data-sift in Old High German data (Luiten et al. 2013). Corpora can provide straightforward information on statistical phenomena in a language such as word and character frequency, which teachers can use to improve their instruction of the language. Lewis (2014) and Wyner (2014) have both suggested language learning approaches that begin with high frequency lexical items, but information about what these are in Cherokee is sorely lacking. Because educating both new second language learners and creating first language speakers is important in revitalizing Cherokee, improvements in pedagogy are crucial. Data-driven approaches to polysynthetic languages with complex morphology are already underway in other indigenous communities, but have not yet begun in Cherokee. Mager et al. have demonstrated the value of this approach in the Uto-Aztecan Wixara, or Huichol language, establishing a parallel (Wixara-Spanish) corpus of Hans Christian Andersen's literature (2018a). Similar corpora of parallel texts are available in Shipibo-konibo, a Panoan language spoken in the Amazon region between Brazil and Peru, as

 ISBN: 978-0-9973295-8-2

well as in Guarani; a member of the Tupi-Guarani family (Mager et al. 2018b). These projects, in addition to serving as the basis for future projects in those languages, increase their visibility in the digital domain.

The current work follows a trend in small language communities using technology for revitalization purposes. Scholars have outlined technology's utility for language revitalization in several ways. Lillehaugen (2017) refers to social media as a means for small languages to reach wider audiences at low costs, and points to the internet as a way for community languages to appear on a global stage. This latter point is vital, considering that these languages are frequently devalued at the local level. In order for threatened languages to persist, it is important for them to establish new domains of use. Otherwise it is far too easy for both speakers and non-speakers to deem these languages to be things of the past (Lillehaugen 2017). More broadly, Crystal (2010: 141) asserts that "[a]n endangered language will progress if its speakers can make use of electronic technology" Part of the reason for this is that the internet can allow speakers of threatened languages to create virtual communities, even when participants are geographically disparate. In essence, the web serves as a vehicle to bring the local to the non-local, as it facilitates communities' capacity for "sharing and interacting with culture, images, and experiences in a small-language context" (Lillehaugen 2017). For Cherokee, the use of technology in revitalization is a natural fit. The language is already included among existing Unicode-compatible fonts, comes standard on all Apple operating systems (Boney 2011), has a Google search page (Cornelius 2012), and has a Facebook translation project underway (Good Voice 2009). This project seeks to expand the existing work on Cherokee revitalization, focusing specifically on the language as spoken in North Carolina.

Documentation has long been understood as a crucial element involved in preserving endangered languages. From the perspective of revitalization, however, documentation has not been enough to assure the language continues to be used in day-to-day life. In fact, very few examples of original Cherokee texts have been recorded in North Carolina – unfortunately much material we have documented exists in translation. At present, Western Carolina University in Cullowhee, NC is in possession of an archive of spoken Eastern Cherokee, but this material has yet to be transcribed.[1] To make documentation truly useful to revitalization efforts, practitioners must be mindful of how the language they document can be applied in returning the language to its community of speakers. Recent technological advancements have facilitated not only the documentation of endangered languages, but the ability to arrange and sift the data such that it will be useful in curriculum development, lexicon creation, machine translation, and much more. This paper articulates the beginning of such a project as leveraged toward revitalizing North Carolina Cherokee among the Eastern Band of Cherokee Indians. I show several ways in which a Cherokee/English corpus can contribute to the contemporary revitalization of Cherokee, rather than simply preserving it for future analysis.

The Eastern Band of Cherokee Indians (EBCI) is located in western North Carolina on land known as the Qualla Boundary. Cherokees ceded most of the land in western North Carolina in an 1817 treaty. The treaty stipulated, however, that the heads of Cherokee families could apply for individual 640-acre reservations, renouncing their citizenship in the Cherokee Nation and becoming citizens of the United States (Finger 1984:10). Consequently, when Cherokees in Tennessee and Georgia were forcibly

[1] Dr. Sara Snyder-Hopkins, Coordinator of Cherokee Language Program, Western Carolina University, personal communication.

removed to Indian Territory (present day Oklahoma) in 1838 along the infamous Trail of Tears, some Cherokees in western NC had a legal basis on which to remain on their ancestral land. In addition to those who stood upon that legal basis, many citizens of the contemporary EBCI draw their ancestry to those who hid in the mountains from U.S. soldiers during the removal, as well as those who returned from Indian Territory after the Trail of Tears had ended.

The Eastern Band shares the Cherokee language with the Cherokee Nation and United Keetoowah Band (UKB), both federally recognized tribal nations headquartered in Tahlequah, OK. Even though the three nations speak the same language, many speakers in North Carolina today, as well as most members of the United Keetoowah Band in Oklahoma, speak the Middle, or Kituwah dialect. Kituwah is one of the three original dialects, alongside the Overhill and Underhill dialects. While Underhill went extinct in the early 1900s, the Overhill dialect continued to be spoken predominantly by Cherokees in Georgia and Tennessee; many of whom were removed to Indian Territory. Because of that, most Cherokee Nation speakers speak Overhill while most NC speakers and members of the UKB speak Kituwah.

Cherokee carries the distinction of having been the first American Indian language to have its own newspaper, the Cherokee Phoenix, which began publication in 1827 – six years after the invention of the Cherokee syllabary (Bender 2002:26). The syllabary is a writing system developed by Sequoyah; a monolingual Cherokee speaker who was previously illiterate in any language. Similar to the function of Japanese *hiragana* and *katakana*, Cherokee's characters each indicate a syllable. The only exception is the ᏍᏓ character, which indicates an [s]. Today, the Sequoian syllabary has been adapted to Unicode and is available as part of all Apple product operating systems as well as Windows and Android. The Cherokee Nation has made significant strides in integrating the syllabary into the fabric of the internet as well, working with Google to establish a Cherokee language verion of the famous search engine. Meanwhile, a Facebook translation project is underway.

Today there are approximately 230 speakers of Cherokee in North Carolina (Micah Swimmer, Adult Language and Education Coordinator, New Kituwah Academy, personal communication), the majority of whom are 65 and older. The language is typically not being passed on intergenerationally in the home. Despite this, there is an immersion school – New Kituwah Academy – that has endeavored to promote the language since 2005. Today New Kituwah extends from preschool through grade six, and children receive their education in the Cherokee language. Because of the large age gap between immersion school students and elders who speak Cherokee as a first language, however, many immersion school students are not exposed to the language beyond school hours. Because students are not hearing or seeing much Cherokee in their day-to-day lives and already speak English as their first language, there is a fear that they will abandon the language for English. To combat this, we must encourage the education of new second language speakers of Cherokee. Because these learners will be acquiring the language as adults, they will have different needs in acquiring the language than children learning it as a first (or child second) language. Among these needs are ample opportunities for practice speaking, listening, reading, and writing in Cherokee. Unfortunately, these opportunities are currently few and far between in the community, and not everyone is informed about those opportunities that do exist.

Although Cherokees have made great progress in making the language usable and available, access issues still remain. Not all first language speakers of Cherokee are

qualified teachers, and second language education in Cherokee is not as strongly informed by current best practices in L2 pedagogy as it is in other world languages. The current project aims to increase available input for language learners, archive existing Cherokee texts in searchable form, and begin iterating on available materials. We can use the texts that exist in the language – from children's books and personal anecdotes to the recent translations of E.B. White's *Charlotte's Web* and Charles Frazier's *Thirteen Moons* – to learn about the Cherokee language and pass that knowledge on to learners. To this end, I have begun a corpus of existing Cherokee language texts and their translations for use in future projects; the utility of which I enumerate below.

Section two of this paper describes the collection of materials that have contributed to the current iteration of the corpus. I provide the names of the particular texts and discuss text types, and lay out the procedure I used in importing the texts to the database. Section three describes what problems can be addressed using corpora, including how they can assist in planning curriculum material and producing new media in the target language. It also speaks to how well-sorted data can answer larger theoretical linguistic questions, such as how polysynthetic languages handle word order given their complex morphological structures. Section four concludes by articulating how the inclusion of corpus materials can help wider language revitalization efforts by leveraging data and creating new tools.

2. COLLECTION OF MATERIALS. The first step in creating this corpus was in locating Cherokee language materials. Through frequent contact with Kylie Crowe Shuler, Bo Lossiah, and Micah Swimmer, administrators at New Kituwah Academy, I was able to amass a collection of texts. Many of these texts were translations of English materials that school faculty and staff had translated into Cherokee, including both popular children's books like *Charlotte's Web* and stories the community members had authored themselves for the school's use (*Buddy the Bluebird* and *The Beast*). That meant that both the English and Cherokee texts were readily available for entry into the corpus. Potentially, it also means there could be discrepancies in the kind of Cherokee the texts represent, as the structures may not be 100% natural in terms of what a speaker might spontaneously produce. Even so, each text was translated from English into Cherokee by an elder who speaks Cherokee as a first language (see Figure 1 for a list of authors and translators). This means that although some texts may come off as stilted Cherokee, there is little probability that they will be expressly ungrammatical. The largest source of data in the corpus so far is the Cherokee translation of E.B. White's *Charlotte's Web*. Future work will add the *Removal* section of Charles Frazier's *Thirteen Moons*, translated by Myrtle Driver Johnson. Other texts included in the corpus's current iteration are children's stories written by EBCI citizens and translated by fluent speakers, as well as a telling of the traditional story *Spearfinger*. For texts that existed in digital form already, it was a simple matter to copy and paste the Cherokee syllabary and English texts each into their own raw text (.txt) file. Some texts required the use of Optical Character Recognition, which exists for Cherokee in rudimentary form via the Tesseract OCR engine.[2] I acquired some texts via a scraper program, which moves text from websites to local hard drive directories. Finally, some of the texts were hand-typed into (.txt) files by Duncan Britton, an enthusiastic undergraduate volunteer.

[2] https://github.com/tesseract-ocr/

TABLE 1. Corpus content as of October 30, 2018.

Title	Author	Translator	Word Count (Cherokee)
Charlotte's Web	E. B. White	Myrtle Driver Johnson	17,913
The Beast	Ben Frey	Marie Junaluska	245
Peas – Our Garden, Our Life	Bill Johnson	Marie Junaluska	162
The Big Journey of Little Fish	Jeffrey H. McCoy	Myrtle Driver Johnson & Abel Catolster	792
Bobby the Bluebird – The Blizzard Blunder	Lynne Lossiah	Myrtle Driver Johnson	308
Spearfinger	Luzene Hill	Nannie Taylor	580
A Very Windy Day	Billie Jo Rich	Myrtle Driver Johnson	108

Once each Cherokee and English text was in its own .txt file, I employed regular expressions (a kind of advanced search and replace feature) to separate each sentence onto its own line within the file. After spot-checking to make sure each sentence was on its own line, I dropped the Cherokee sentences into a single column in an Excel spreadsheet with the English sentences in the column beside it. I then read through each sentence pair to check whether the sentences truly corresponded. In some cases the English or Cherokee text was longer. Often this represented material lost or gained in translation – some idiomatic expressions in one language or the other do not translate succinctly and sentences had to be added or subtracted. I found the most efficacious way to solve the problem was to combine two English or Cherokee sentences onto the same line in the Excel spreadsheet beside the single sentence to which they corresponded in the other language.

After assembling alignment files in Excel, I dropped the aligned Cherokee sentences back into a .txt file and dropped the English sentences into a separate one. I used AntPConc (Anthony 2017) to designate the English .txt file as the English corpus and the Cherokee .txt file as the Cherokee one. Doing so made it possible to query the database in either language to search for individual English or Cherokee words. I was also able to designate the Cherokee .txt file alone as its own corpus in AntConc (Anthony 2018), which allowed for word and syllabary character frequency counts. Frequency counts will facilitate second language acquisition, allowing teachers to focus first on the most frequently-occurring words and characters. This will allow students a feeling of having "easy wins" early on, as well as deriving the greatest benefit from

some of the earliest forms learned. Acquiring high frequency words and characters first can reduce the difficulty curve in learning a second language (Ferris 2012).

3. SOLVING PROBLEMS WITH CORPORA. Scholars in SLA research have illustrated the usefulness of corpora in the language classroom. Teachers can, for instance, use a corpus of interactions within certain event types (meetings, presentations, discussions over coffee, etc.) to help students learn what expressions may be useful for certain communicative functions (i.e. expressing disagreement, asking questions, etc.) (Mauranen 2004). Corpora, and in particular frequency and distributional information, can also reveal information about the semantic, discourse, and syntactic contexts in which words occur – information that cannot be found in dictionaries or grammars (Pereira 2004). Particularly useful is the analysis of "chunks" of language; sets of words that co-occur on a regular basis ("I mean," "this that and the other," etc.) (O'Keefe et al. 2009). For Cherokee, corpora can help in two key ways: they can help to improve language teaching pedagogy and supply more Cherokee reading and teaching materials. Much of the pedagogy for teaching Cherokee in North Carolina until recently has consisted of first language speakers listing words and phrases on a white board and having students copy them, with occasional pronunciation practice. One of the major goals of my research is to improve on these pedagogical techniques in order to increase the number of proficient second language learners. This goal arises largely from my earliest efforts to learn the language as a citizen of the Eastern Band of Cherokee Indians, beginning in 2003. I want to facilitate efforts at language learning for other tribal citizens as well as for non-Cherokees because my own efforts were so trying.

Research in the field of Second Language Acquisition (SLA) on Communicative Language Teaching (Omaggio Hadley 1993; Lightbown & Spada 2013), can help parlay theoretical linguistic information into more effective pedagogy. Reference to existing textual materials can provide insight into the usage of particular vocabulary items, the collocations of verbs, and the general structure of Cherokee sentences. While these domains are well-articulated in the linguistic literature, they are under-utilized in teaching contexts. Proficient adolescent and adult second language speakers will, in turn, be able to support young learners and carry the language beyond the borders of the immersion school and into the community. In order to do this, however, we need to establish a link between theoretical linguistic research and good SLA pedagogy. Linguists have long focused on documenting endangered languages and analyzing their structure. Their hope has been to contribute to the pool of human knowledge on how languages function in general, yet a different tack may contribute to pulling these languages back from the brink. By abstracting linguistic patterns into learnable rules, second language learners may become proficient speakers of languages that are currently endangered or even dormant; potentially leading to fluent first language speakers in the following generation. Even in the absence of discrete *rules* for language learning, being able to model language lessons on real language will be crucial to second language teachers – students may not learn rules overtly, but will be able to infer them from exposure to accurate examples. This is where corpora can be extremely valuable.

Because many texts in the corpus are already translations of English texts, they serve as a good model for what structures are acceptable to translate from English into Cherokee. While translation from English may not produce the most representative samples of Cherokee language, they benefit from the ubiquity of existing English

materials. Translation is a quick route to a high volume of reading, viewing, and listening materials in Cherokee, and also allows the tribe to exploit the existing popularity of certain English language characters and stories. The creation of a corpus can facilitate quicker, more accurate, and more streamlined translation from Enlgish to Cherokee. Because second language learners and immersion school students are in constant need of new reading material in Cherokee, a demand exists for both texts originally written in Cherokee and translated texts. It took Myrtle Driver Johnson, Cherokee Beloved Woman and fluent speaker, 3 full years to translate *Charlotte's Web* from English into Cherokee. A corpus of texts can provide the basis for creating Computer-Assisted Translation (CAT) tools, and ultimately even training data for a neural network such as the one used by Google Translate. Such tools can never replace a fluent speaker, but could assist them in their considerable tasks. By crowd-sourcing the initial pass of a translation with CAT tools, Cherokee language students might create a rough translation that could then be proofread by teachers and fluent speakers. This should reduce the workload for fluent speakers and make the task of translation slightly less daunting. While CAT tools are not particularly well-suited to literary translation, they could prove invaluable in generating largely fact-driven and/or repetitious texts such as documentaries, manuals, restaurant menus, grocery store item labels, etc. After enough material has been added to the corpus, it will serve as useful training material for machine learning systems. Similarly, a corpus of spoken Cherokee would facilitate the creation of speech recognition and speech-to-text tools that could further aid in revitalization attempts. For learners, text-to-speech engines trained on a corpus of spoken and written texts would be useful in many formats – from producing examples for dictionary entries to use in second language learning software.

Compiling a textual corpus will also facilitate the creation of dictionaries. Rather than entering words one by one, Cherokee lexicographers could reference words within a corpus, providing not only a definition but also a contextualized example sentence. Assuming the corpus contained a broad enough array of genres, word-frequency lists generated from a corpus would also inform second language teachers about what words would be most productive to teach beginning students. Table 2 shows a sample word frequency set derived from the current corpus.

TABLE 2: Forty most frequently occurring words in Cherokee corpus as of November 18, 2018.[3]

Syllabary	Roman orthography	English gloss	Number of Occurrences
Ꮓᏸ	Nole	and	789
ᎤᏛᏁ	Udvne	(s)he said	434
ᏴᏍᏓ	Gesdi	not	308
ᎣᏍᏏ	Osda	good	211
ᏯᎩ	Yigi	if it is	139
D4Z	Aseno	but/however	135
ᎪᎲᏍᏓ	Gohusdi	something	98
ᎢᏳᏍᏓ	Iyusdi	like/as (similar to)	96

[3] Frequency list curated to omit genre-specific vocabulary items like personal names.

ᏂᏚᏓ	Nigada	all/everyone	93
ᏃᏗ	Nodi	now then	92
ᎢᏍ	Iga	day	88
ᎤᏍᏗ	Usdi	small/baby	82
ᎢᏣ	Itsa	toward	78
ᏱᏎ	Gese	it was (non-evidential)	73
ᎤᎷᎷᏁ	Utvdvne	(s)he asked	72
ᏧᏂᏍᏆᏂᎪᏙᏗ	Tsunisquanigododi	enclosure	66
ᏰᏢ	yeli	if it is possible	66
ᏱᏒ	Gesv	it was (evidential)	65
ᏍᎩᎾᎾ	Sginana	and thus/and then…	63
ᏐᏉᏫ	Soquo	one	63
ᎡᎳᏗ	Eladi	low	62
ᎾᎥᏂ	Navni	near it	62
ᏍᏉᏫ	Squo	too/also	60
ᏱᎦᎢ	Gesei	it was (non-evidential); full form	58
ᏱᎦᏍᏗ	Gesesdi	it will be	56
ᏯᏀ	Kilo	someone	55
ᏑᎾᏧ	Sunale	morning/tomorrow	53
ᎡᏝᏯᎯ	Etlawehi	Quiet, silence	51
ᏣᏍᏗ	Tsesdi	Stop it!	51
ᎠᏍ	aya	I/me	50
ᏃᏉᏫ	noquo	now	50
ᎠᎬᏰ	agvyi	first	49
ᏫᏁᏆᎢ	kanesai	box	48
ᏧᏯᏂᏐ	tsuwetsi	his/her egg/child	48
ᎠᏥ	ama	water	47
ᏰᏱᎡᎾ	yigesvna	without doing it	47
ᎢᎪᎯᏓ	igohida	a duration; until	46
ᎯᎠ	hia	this	46
ᏩᏀ	hawa	alright, okay	44

With a very robust sampling of texts, a corpus would approach representation of the language at large, providing true insight about what words occurred most frequently in a general sense. Students could make use of that statistical knowledge in order to make great initial strides in language learning. A sampling of various genres would also enable researchers to generalize about the features of particular textual genres and how they are constituted in Cherokee. Even if the corpus under consideration were not representative of the language as a whole, teachers could key their lessons toward particular texts or text types they wanted students to learn, mining vocabulary lists from the corpus that were relevant to those particular texts. In designing a lesson on traditional stories, for example, a teacher might select words that occurred with high frequency within those stories. This could generate a vocabulary list for students to

study before reading the story, and form the basis for lessons about those vocabulary items. By priming students before reading, teachers facilitate students’ comprehension of the input, making it more likely students will understand and retain the language (Krashen & Terrell 1983). Frequency of occurrence is also extremely useful in targeting the acquisition of the Cherokee syllabary – because syllabary characters do not occur with uniform frequency, teachers could lighten students’ mental loads by focusing initially on high frequency characters rather than simply teaching the syllabary by rows. Table 3 shows a breakdown of syllabary character frequency in a recent data set.

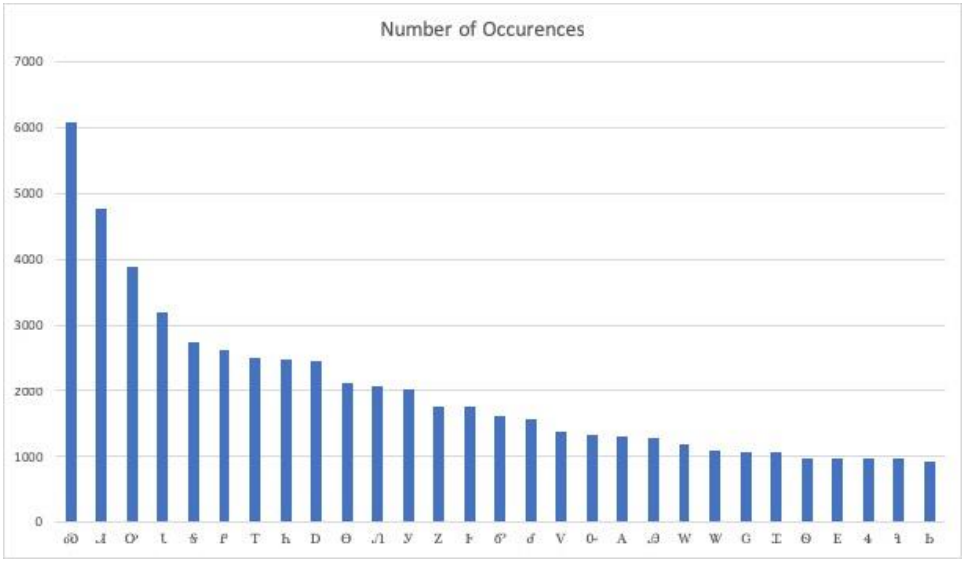

FIGURE 1. Thirty most commonly-occurring syllabary characters for corpus as of November 4, 2018.

Pilot projects using CAT tools for translating texts from English to Cherokee are already underway. One such project is the translation of the open-source match three puzzle game *Heriswap*, available on the Google Play Store. After surveying available open-source projects, Dr. Derek Lackaff (Elon University) and I produced a Cherokee language version of the game shown in Figure 2. By increasing the availability of games, apps, and software tools in the Cherokee language, we hope to encourage broader use of the language both by community members and second language learners. Although our process in creating the translation of *Heriswap* was to extract the strings needed and translate them one by one, a corpus of Cherokee texts could provide some “bootstrapping” of such projects via a consistently maintained translation memory (.tm) file that would log existing translations of strings and be able to suggest them when they reoccur in other translation

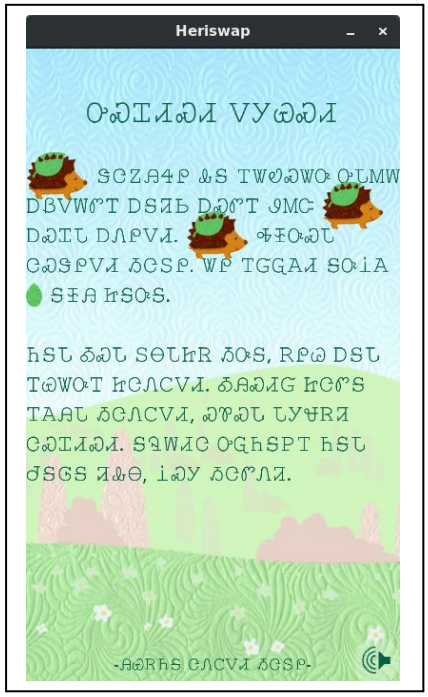

FIGURE 2. Cherokee translation of *Heriswap*.

projects. For example, there would be no need to translate the string "start game" for each new game translation project, assuming that phrase had been translated once and stored in a .tm file for future translators' use. A CAT tool such as OmegaT or SDL Trados would simply suggest the existing translation, and translators could use their discretion in deciding whether they wanted to use it or not.

Like any translation project, participating in translation projects of games and apps would provide opportunities for second language learners to polish their Cherokee language skills. Their "verified translations" could be passed on to teachers to be proofed, and teachers could forward these to native speakers in order to assure accuracy. Once translations had been approved, they could be fed back into the corpus to provide more data for future projects. This would create a virtuous cycle, easing translation while putting second language learners and Cherokee language teachers in close collaboration with native speakers from the Cherokee communities.

Text translation and creation of new Cherokee language materials help stem the tide of English dominance in society at large. One key factor in driving language shift is the exposure people have to one language over another in their day-to-day lives (Frey 2013). Because we live in a society constantly connected to the internet and surrounded by media in various forms, it is extremely important that we be able to experience that content in the language we wish to revitalize. If people do not speak, hear, read, and write the language on a regular basis, their facility with it will only continue to decline. Having children's books is therefore vital, but also not enough. If we are to truly see a reversal of the shift toward Anglo-centrism, we must take steps to reduce its overwhelming presence in our communities in favor of our own language. The best way to do that is for the language to be transferred from generation to generation in the home, but for many people in the Eastern Cherokee community that option no longer exists. That is why we must scaffold language learning opportunities with ubiquitous opportunities for exposure to Cherokee in day-to-day life.

Even if second language learners are not able to work closely with fluent speakers, a corpus can provide access to speaker-generated materials that can guide the acquisition process. Provided that texts in the corpus were produced by native speakers, the texts' grammatical constructions will represent accurate Cherokee forms. On that basis, teachers could help students to "mine" sentences from the texts that contain forms students might want to learn and use those to create activities for classroom use. This would be an easy source of material for flash cards and "gap fill" activities, in which students must fill in a blank with the correct word. A corpus would provide a nearly endless supply of example forms, which teachers could integrate into such exercises. Teachers and students could also begin extrapolating on the structure of example sentences and substituting in different words to make their own, grammatically accurate, parallel sentences. A rudimentary example would be, upon finding a sentence like "The dog ate all of his food," a class could turn that sentence into "the *cat* ate all of *my* food." Substituting the word "dog" for "cat" and "his" for "my" is trivial, but it substantially changes the meaning of the sentence in systematic ways that students can follow. Exercises like this provide not only grammatical scaffolding and understanding, but can also be sources of humor and language play. Figure 5, below, presents the first four results that appear in the database when querying the English word "ate."

TABLE 3. Sample results of a query of the English “ate.” [4]

1.	While Wilbur	ate,	Lurvy fetched a hammer and some 8-penny nails and nailed the board in place.
	ᎤᏁᏍᎳᏴᏓ ᏕᏒᏙᏓ ᎩᎷᏤᎢ, ᎤᏓᏔᏍ ᏕᏁᏯ, ᎤᏲᏏᏍᎩ ᎤᏓᏔᏍ ᎤᏅᏓ ᏃᎴ ᎤᎩᏍᏙᏒ		
	ᏯᎾᎴ ᎠᎩᏍᏓ ᎤᏂᏯᏀᏏ. ᏏᏍᏓ ᏣᏁᏍᎳᏁᏍ ᎤᏩᏂ ᎤᎩᎩᎤᏒ ᎡᏙᏯ. ᏕᎤᏆᏩᏍᏓ ᎩᏴፍ ᏣᏀᏏ		
	ᏃᎴ ᏣᏁᎳ ᎠᏂ4ᏂᏏ ᏴᎩ, ᏏᏍᏓ ᏣᏩᏁᎴ ᎠᏦᏴ.		
	Ulistayvdi galidodi wulutsei, uditahe ganehv, uyosisgi uditahe unvdi nole ugisdoe sunale agisdi unisulvda. Osda nulistanehe uwasa uwenvsv edohv. Ganvqualosdi wuyvhe tsanusi nole tsanela anisenisi yvgi, osda nuwanele atsoyv.		
2.	Mr. Zuckerman sat down weakly and	ate	a doughnut.
	ᎤᏁᏍᎩ ᎢᏳᏍᎩ ᎤᏁᏍᎤᎴᏁᎢ ᎭᎻ ᏃᎴ ᎤᏕᎾᏍᏓ ᏕᏚ ᎠᎳᏣᎩ ᎤᏂᎢ.		
	Ulisdi iyusdi ulisdvtsunei homi nole uganasda gadu atalvgidi ugei.		
3.	Charlotte rested and	ate	a grasshopper.
	ᏌᎳᏓ ᎤᏣᏘᏥᏍᏓᏕᎢ ᏃᎴ ᏙᎱᏓᏍᎥ ᎤᏩᏍᏂᎢ.		
	Salada utsawesolvsdei nole dolisqua uwayagei.		
4.	She	ate	a small bug that she had been saving.
	ᎤᏍᎩ ᏍᎪᏍᏃ ᎤᏍᎥᏂᎪᎾᏛ ᎤᏩᏍᏂᎢ.		
	Usdi sgoya usquanigodv uwayagei.		

The first example does not match the English text one-to-one. Instead, the translator, Myrtle Driver Johnson, paraphrased the English in order to translate it into Cheroke. Rather than translating the English into one sentence in Cherokee, she divided it into a few sentences in Cherokee. Additionally, all of these examples demonstrate the breadth of options a Cherokee speaker has to translate the word “ate.” In order to translate from English, Johnson uses the word ᎤᎩᏍᏙᏒ (*ugisdoe*) in the first sentence – a word indicating that a food was chewed. The first clause of the English sentence, “While Wilbur ate,” is not translated as part of the Cherokee sentence “ᏕᎤᏆᏩᏍᏓ ᎩᏴፍ ᏣᏀᏏ ᏃᎴ ᎠᏂ4ᏂᏏ ᏴᎩ, ᏏᏍᏓ ᏣᏩᏁᎴ ᎠᏦᏴ.” As it stands, the Cherokee sentence would translate back into English as “Lurvy fetched a hammer and some eight-cent nails and repaired the fence.” It maintains the spirit of the English text, but does not preserve every detail. Instead, Wilbur’s eating is referred to two sentences prior. The second sentence uses a fairly straightforward word for “to eat,” ᎤᏂᎢ, built on the root -gi-, “to eat a solid object,” while the next two sentences classify the shape of the object being eaten (an insect) as flexible, ᎤᏩᏍᏂᎢ (*uwayagei*).

4 Data from ᏌᎳᏓ ᏚᏏᎳᏛ, *Charlotte’s Web* by E.B. White, translated into Cherokee by Myrtle Driver Johnson. 2015 Eastern Band of Cherokee Indians.

From this small sample, we can see that Cherokee has a range of options available to translate English terms, especially when we consider the 5-way classifier system that divides direct objects into solid, liquid, flexible, rigid, and animate shape categories. These sentences, and the particular lexical items used, can prove useful for students if we maintain the senses in which each verb is used. From the above, we know that the verb ᎤᎨᎢ can refer to a solid object like a doughnut. We should be able to extend that to other objects, as long as they are similarly shaped. Hence, we could imagine Mr. Zuckerman eating an apple instead of a doughnut, and write this sentence simply by substituting the word “doughnut” (ᎤᎦᎾᏍᏓ ᎦᏚ ᎠᏔᎸᎩᏗ - *uganasda gadu atalvgidi*) for the word “apple” (ᏒᎦᏔ - *svgata*):

“ᎤᎵᏍᏗ ᎢᏳᏍᏗ ᎤᎵᏍᏛᏧᏁᎢ ᎰᎻ ᏃᎴ **[ᎤᎦᎾᏍᏓ ᎦᏚ ᎠᏔᎸᎩᏗ]** ᎤᎨᎢ”
“ulisdi iyusdi ulisdvtsunei homi nole [**uganasda gadu atalvgidi**] ugei” ➔
Mr. Zuckerman sat down weakly and ate [**a doughnut**].

“ᎤᎵᏍᏗ ᎢᏳᏍᏗ ᎤᎵᏍᏛᏧᏁᎢ ᎰᎻ ᏃᎴ **[ᏒᎦᏔ]** ᎤᎨᎢ.”
“ulisdi iyusdi ulisdvtsunei homi nole **[svgata]** ugei.”

Although replacing a noun in the text for another seems trivial, using samples of existing texts reveals complexities we might not otherwise account for. Although learners might assume that “ate” could take any direct object based on the English verb’s meaning, looking at the translations reveals parameters, like shape classification, the learner might not have considered. Based on this example, we can begin experimenting with solid objects characters might eat as well as contrasting our sentences with ones that refer to eating flexible objects. Instead of a bug, we might talk of a person eating a well-cooked steak. It should be noted that although the first four results from the database come from the text *Charlotte’s Web*, querying the word “ate” returns 39 results from a range of texts.

Sentence mining techniques like this also provide a window into the general structure of Cherokee word order. This is particularly important from a theoretical standpoint. Montgomery-Anderson (2008: 25) notes that “[t]he current literature is … lacking many details of the syntax of the language” while Beghelli (1996: 105) characterizes Cherokee syntax as “largely unexplored territory”. Existing scholarship has posited that word order in Cherokee is either free or governed by a principle of “newsworthiness,” (Scancarelli 1987; Mithun 1987; Montgomery-Anderson 2010, 2016) but has not provided a general rule of thumb for students to follow when ordering Cherokee sentences. Indeed, Montgomery-Anderson (2008: 115) observes that “[t]he idea of ‘basic’ word order is problematic in Cherokee. While there are word orders that are more common than others, it appears that, given the right context, most word orders are possible”. This corpus will allow us to directly probe the idea of a ‘basic’ word order, and, in the absence of such a phenomenon, to generalize about when particular orders occur. Most studies of Cherokee grammar to this point have, with good reason, focused on its complex morphology. Once learners of the language begin to get a handle on verb conjugation, however, they will need more robust phrase structure rules in order to both interpret and create novel Cherokee sentences. A large collection of existing Cherokee sentences will allow us to ascertain the practical distribution of word order in the language. Although theoretically, Cherokee’s morphology should *allow* a generally free word order, a corpus can help discover what speakers and/or authors

actually *do* when creating texts. If, for example, we can ascertain that 70% of sentences are Subject-Object-Verb (SOV) or VOS, we could provide that as a general template on which students could build their sentences. Further research could then discover why deviations existed and what conditioned them.

This kind of large-scale data sifting also has applications for theoretical research. A thorough sort of the data would allow inquiry into broader patterns. Linguists know, for example, that Cherokee attaches a relativizer prefix *tsi-* to create relative clauses, but what is the prevalence of that relativizer in comparison with wh-question words? What kinds of words or structures condition a change in verb stem, and how many instances of each verb stem can be contained within a sentence? Collection of a corpus, along with thorough part of speech and morphological tagging, can provide insight into such questions. This, in turn, would yield further information for students of the language.

One way to streamline second language learning is to gear initial lessons toward the most commonly-occurring words in the language. Assuming a corpus was broad enough, it could generate a list of most frequently-occurring words that was representative of the language at large. For Cherokee, the concept of words vs. phrases is somewhat problematized due to the language's complex morphology. The solution I propose would be to query which particular verb forms occur most frequently and extrapolate from that which forms would be most helpful to teach. If the form *hega*, "you are going," occurs in the list of high frequency words, for example, instructors could opt to teach it as well as forms like *uwenvsdi*, "for him/her to go." The operative piece of information would be that the verb "go" is frequently occurring, and teaching its five stems (Montgomery-Anderson 2016) would therefore be useful to second language learners. By learning the 1,000 or so most frequently-occurring words (or forms) in a language, a learner should be able to understand as much as 85% of daily conversations (Lewis 2014). By learning the 3,000 most common words, that percentage may increase to 95%. Of course, learning words in context is also crucial, and a corpus can provide thousands of examples. Second language teachers could create gap texts out of sentences from the corpus, focusing on a particular word or construction they wanted students to attend to. This would reduce the workload of teachers by alleviating the need for them to think of dozens or even hundreds of novel example sentences. Students would also benefit from seeing the language as it has truly been used in texts, rather than simply being given a list of prescriptive rules or intuitive judgments about how speakers suspect the language should be. By basing their language on how speakers have used the language in the past, students may approach a more accurate spoken and written Cherokee than they otherwise would have.

4. PITFALLS. The methodology presented here of using parallel (English/Cherokee) texts is not perfect. Even though the translators of the English texts are first language speakers of Cherokee, there is potential for structural overlap between the two languages. Future work will transcribe and integrate material from Western Carolina University's archive of spoken Eastern Cherokee into the database, as free conversation will yield more relevant and useful results. The corpus at present also suffers from being too small, lacking sufficient variation in text types to accurately reflect the breadth and depth of the language writ large. The inclusion of more and longer texts such as the Eastern Cherokee translation of *Encyclopedia Brown* and the "Removal" section of Charles Frazier's *Thirteen Moons* should help to balance the corpus, as will the inclusion of transcribed spoken material. Future work will also collect new spoken

narratives from elders in order to include more original Cherokee language material. The present work is intended as a proof of concept, illustrating the initial steps of how this work can be done. It is my hope that other scholars of Cherokee will see these initial attempts and be inspired to collaborate in a fashion befitting the value of ᎦᏚᎩ (*gadugi*) – people coming together as one and working to help one another. As a citizen of the Eastern Band, it is my intention to work in community with other Cherokee people, speakers, and scholars to further the goal of revitalizing the language. I see these initial steps as a contribution, though I am aware they do not constitute a finished product.

5. CONCLUSION. No single technological solution will save endangered languages. For that we have to begin using them in our communities in day-to-day life and passing them on to future generations as a first language. Even so, many lack the skills to even begin this journey, and technological solutions may facilitate that beginning. By integrating corpora into second language learning and language revitalization efforts in general, we can accomplish a great deal more than we would have without these tools. Corpora can inform second language teaching pedagogies, provide a basis for rapid translation of materials, and provide useful theoretical insights about the general structure of the language under discussion. Corpora also serve the greater good of the language by recording it in its current state for posterity. Regardless of the outcome of our language revitalization efforts today, collecting a corpus of natural language from fluent speakers and authors will ensure that the language carries on into the future in one form or another. This way, even if the language becomes dormant at some point in the future, it will still be able to be brought back through reference to the texts in the corpus. My own forays into corpus building are in their infancy, but I hope to be able to realize many of the possibilities I have articulated here, and share those possibilities with other scholars struggling with the issues of language endangerment and revitalization.

REFERENCES

Anthony, L. 2017. AntPConc (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Anthony, L. 2018. AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Beghelli, Filippo. 1996. Cherokee clause structure. In Munro, Pamela. (ed.). Cherokee Papers from UCLA-UCLA *Occasional Papers in Linguistics 16*, pp. 105-114. Los Angeles: University of California.

Bender, Margaret. 2002. *Signs of Cherokee Culture: Sequoyah's Syllabary in Eastern Cherokee Life*. Chapel Hill: University of North Carolina Press.

Boney, Roy Jr. 2011. The indomitable language: How the Cherokee syllabary went from parchment to iPad. *Indian Country Today*. 1(35). Retracted from: https://conceptualizingtuscarora.wordpress.com/2011/10/03/the-indomitable-language/ (Data accessed: November 4, 2018).

Cornelius, Craig. 2012. "ᎯᏧᎾ ᏗᏓᏴᎳᏛᏍᎩ ᎬᏗ Gmail ᏣᎳᎩ (Get started with Gmail in Cherokee." Official Gmail Blog. https://gmail.googleblog.com/2012/11/gmail-get-started-with-gmail-in-cherokee.html (Data accessed: November 4, 2018).

Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.
Ferris, Timothy. 2012. *The 4-Hour Chef: The Simple Path to Cooking Like a Pro, Learning Anything, and Living a Good Life*. New Harvest.
Finger, John R. 1984. *The Eastern Band Of Cherokees, 1819-1900*. Knoxville: University of Tennessee Press.
Frey, Benjamin. 2013. *Toward a General Theory of Language Shift: A case study in Wisconsin German and North Carolina Cherokee.* Ph.D. dissertation, University of Wisconsin-Madison.
Good Voice, Christina. 2009. Cherokee language now on Facebook. *Cherokee Phoenix*. https://www.cherokeephoenix.org/Article/index/3513 (Data accessed: November 4, 2018).
Krashen, Stephen D. & Tracy D. Terrell. 1983. *The Natural Approach: Language Acquisition in the Classroom*. San Francisco, CA: Alemany Press.
Lewis, Brendan Richard. 2014. *Fluent in 3 Months: How Anyone at Any Age Can Learn to Speak Any Language from Anywhere in the World*. New York, NY: HarperCollins.
Lillehaugen, Brook Danielle. 2019. Tweeting in Zapotec: social media as a tool for language activists. In Gómez Menjívar, Jennifer Carolina & Gloria E. Chacón (eds.) *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*, pp. 201-226. Tucson: University of Arizona Press.
Lightbown, Patsy, and Nina Spada. 2013. *How Languages Are Learned*. 4th edition. Oxford: Oxford University Press.
Luiten, Tyler, Andrea Menz, Angela Bagwell, Benjamin Frey, John Lindner, Mike Olson, Kristin Speth & Joseph Salmons. 2013. Beyond the handbooks: a quantitative approach to analysis of Old High German phonology and morphology. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)*. 135.1-18.
Mager, Manuel, Diónico Carrillo, & Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for the Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5): 3081–3087.
Mager, Manuel, Ximena Gutierrez-Vasques, Gerardo Sierra, & Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the Americas. *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 55-69. Santa Fé, NM.
Mauranen, Anna. 2004. Spoken corpus for an ordinary learner. In: Sinclair, John McH. (ed.), *How to use corpora in language teaching*. John Benjamins: Amsterdam/Philadelphia.
Mithun. Marianne. 1987. Is basic word order universal? In Tomlin, Russell (ed.), *Grounding and coherence and in discourse*. Typological Studies in Language 11, pp. 281-328. Amsterdam: John Benjamins.
Montgomery-Anderson, Brad. 2008. *A Reference Grammar of Oklahoma Cherokee*. Lawrence, KS: University of Kansas dissertation.
Montgomery-Anderson, Brad. 2016. *Cherokee Reference Grammar*. Norman, OK: Oklahoma University Press.
O'Keefe, Anne, Michael McCarthy, and Ronald Carter. 2009. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge, United Kingdom: Cambridge University Press.

Omaggio Hadley, Alice. 1993. *Teaching Language in Context*. Boston, MA: Heinle & Heinle.

Pereira, Luísa Alice Santos. 2004. The use of concordancing in the teaching of Portuguese. In: Sinclair, John McH. (ed.), *How to use corpora in language teaching*, pp. 109-124. Amsterdam/Philadelphia: John Benjamins.

Reppen, Randi. 2010. *Cambridge Language Education Series: Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Scancarelli, Janine. 1987. *Grammatical relations and verb agreement in Cherokee*. Los Angeles, CA: UCLA dissertation.

Wyner, Gabriel. 2014. *Fluent Forever: How to Learn Any Language Fast and Never Forget It*. New York, NY: Harmony Books.

Benjamin Frey
benfrey@email.unc.edu