UNIVERSITY of LIMERICK

OLLSCOIL LUIMNIGH

Natural Language Processing and the Mohawk Language

Creating a finite state morphological parser of Mohawk formal nouns

Alicia Alexandra Assini

MSc candidate in Multilingual Computing and Localisation 2012/2013, CSIS

Advised by Richard Sutcliffe, PhD

11/10/2013

Presented in this thesis is the design, implementation and evaluation of a finite state morphological parser for Mohawk formal nouns. Utilizing the finite state morphology software designed by Beesely and Karttunen (2003) along with three of the most comprehensive grammars for Mohawk, one from each of the major dialectal regions, a lexicon for a finite state system was created that incorporated a structure I created from cross-referencing the three sources. Since there was no formal coursework in the program providing instruction in computer programming or morphology, these skills were self-taught. In addition to the parser, a taxonomy of Mohawk prefixes and suffixes was developed for the finite state system. The challenges facing the development of Natural Language Processing tools and language-learning technologies for the Mohawk language, as a polysynthetic language are representative of the problems that other related languages in this important group experience. The research shown here demonstrates that using finite state morphology techniques that the Mohawk language's formal nouns can be successfully described and parsed and that with further study and review that it is possible a finite state system could be expanded to include all of Mohawk nouns, verbs, and particles. This work is looking to provide greater access to the Mohawk language for the speakers and learners of the language by use in an e-dictionary and also to support projects for language tools development for other polysynthetic and under-resourced languages.

Table of Contents

Figures and Tables

Figures:

Tables:

Summary

The goal of this research was to create a morphological parser for Mohawk formal nouns using finite state morphology. The Mohawk language is a member of the Iroquois Language Family, which is a language family native to North America. The Iroquois language family and the people who speak these languages have a long historical and cultural involvement in both colonial American society and in the formation and history of United States of America. Additionally, from a linguistic perspective these polysynthetic languages are valuable to study due to their typology and radically different structure from the more researched European languages and even East Asian languages. Applying various approaches to create morphological parsers for these languages, such as finite state morphology, tests the limits of what languages these approaches and methodologies can be used for and informs researchers of the unique challenges that these languages pose to Natural Language Processing and linguistic analysis.

Using the Xerox Finite State Morphology software a morphological analyzer was designed and created using the lexc programming language to handle Mohawk formal nouns including varying affixes, attributive, locative, and possessive, that can be applied to the noun roots. The analyzer is meant to be a system that demonstrates that finite state morphology can be utilized for this part-of-speech for the Mohawk language. Thirty noun roots were chosen for this system in order to be examples of the varying formal nouns that would take each of the different pronominal prefixes and suffixes, with the goal such that a future researcher could build upon this program and add additional noun roots and suffixes without having to radically alter the structure of the program. Around these thirty noun roots a variety of affixes was included. The system was tested by inputting various Mohawk noun roots with the varying affixes both attempting legal Mohawk combinations and illegal combinations to test that the system could give a correct output when a legal combination was input and would have no output when an illegal Mohawk word was input. The system was successful when a formal noun in Mohawk was the input. With only thirty noun roots, nominal and possessive prefixes and sixteen possible suffixes almost 7,000 words can be created from just thirty noun roots, most of which are legal combinations in Mohawk except when an animate object is paired with an inanimate plural suffix and vice versa, which is only in a few instances. Overall, the output is the correct noun root, or the English translation, including what type of affixes and the meaning of the affixes surrounding the noun root. The program can also take an English word as an input and give an output in Mohawk that includes the word in English being converted to a noun root and the nominal affixes being added to make it a legal Mohawk word.

The design of this system demonstrates that it is possible to create a finite state system that describes Mohawk formal nouns. However, more work still needs to be conducted even within the formal nouns program before beginning work on Mohawk verbs, which is where the majority of the

complexity and content of Mohawk lies. These next steps must be accomplished in order for Natural Language Processing tools and support to become available in the Mohawk language.

1. Introduction

1.1 Natural Language Processing

Technology's continued growth and innovative development has been fast-paced and its impact far-reaching. However, developments involving human speech recognition have not kept up at the same pace and have moved slower than predicted in the 1968 movie *2001: A Space Odyssey*. In 1968 when the film was released it was envisioned that by 2001 scientists would have solved the problem of computers comprehending and recognizing human speech. The film showed that the creators of *2001: A Space Odyssey* foresaw a future where scientists would have figured out how to create interactive computers that speak English and can respond in way that resembles human speech. Indeed, progress has been made on this front of human computer language communication and understanding, but in practice these tasks have proven very challenging for researchers and linguistics to overcome, and there still is nothing quite like H.A.L. 9000 available in 2013.

However, while H.A.L. 9000 won't be available by Christmas 2013, current needs and uses of technology already incorporate many advances and innovations developed by computational linguistic researchers. The study of how computers interact with and process human speech and language is often referred to as Natural Language Processing (NLP) and innovations in this field help billions of people use computers, cell phones and smartphones, and search the internet each day. Some of the products created in the field of NLP include syntactical parsers, morphological parsers, Part-of-Speech (POS) taggers, word sense disambiguation programs, information retrieval applications, speech recognition processing, spell checkers, and machine translation engines. While it goes unnoticed to the average user of the phone or computer, advances in NLP are involved in search engines, and automated telephone operators with speech recognition. Additionally, many young people currently entering the work force grew up writing essays and reports for elementary school on a computer with a word processing program that had spelling correction capabilities enabled, which is another very common product of NLP research.

With the development of computers and the internet, the first software programs primarily were only available in English, as most of the developments took place in the United States and as such the first NLP programs were created for English with research for Machine Translation during the 1960s. NLP involves a series of stages of linguistic processing that can be organized in to a pipeline and then run on a document. One example of a recent NLP and language engineering program is GATE, which was developed in the UK, and allows linguists to perform varying syntactical analysis on English language documents and the creation of pipelines streamlined for the analysis of the users deciding (Haribhakta and Kalamkar et al., 2012, pp. 308—313). Other computational linguists have since developed programs that work within GATE, and are called plug-ins, to provide support for linguistic analysis on documents in other languages and which can be used in the creation of pipelines. Gate is a computer program that most students studying NLP are exposed to when they learn language engineering techniques and approaches to POS tagging, parsing, and spell checking. Morphological analysis is another task that can be inserted in to the NLP pipeline and for polysynthetic languages is in fact one of the most basic of needs for NLP support of a language. However, the creation of a morphological parser for polysynthetic languages is not only more complicated than the more analytical languages but also faces many challenges due to lack of resources because there has not been much study of many of the world's polysynthetic languages. A morphological parser performs a morphological analysis of each word determining the root or stem of the word, and the meaning of the affixes attached to that root. Once this tool has been developed then POS tagging, spell-checkers and information and retrieval systems can be designed around that initial morphological analyzer. GATE is a great program to learn and experiment with different types of linguistic analysis, the order of application of these tools to a language document, and for understanding the unique role of each tool.

1.2 NLP for the World's languages

At this point and time there is NLP support for many of the Western European languages and for the most mainstream Asian languages, such as Japanese, Korean, and Chinese. It is important to add that there are almost 7,000 languages spoken in the world today, and that even though companies

like Microsoft have increased efforts to provide linguistic support of their products to a wider market, they still can only manage to offer their products and services in limited capabilities for around 100 languages. The reality is that people are relying more and more on technology and having increased access to the internet. If people cannot navigate this technology in their own language there are two results: they will remain at a disadvantage since they will not be able to access this information, and they will move away from their native language to one that allows them to access this global network.

What is more, the development of language resources for technology has not been chosen because of the number of speakers but representation is disproportionately high toward European languages, some of which have less than 1 million speakers, such as Estonian. All of this while languages in Africa or India that have well over 70 million speakers such as Twi for example, which is a language spoken in West Africa, have almost no computational linguist support nor online presence (Ethnologue, 2013). It is easy to criticize the development of this situation but some of it is understandable as it is based on the historical ties between the United States and Europe culturally and in business, and also by the business drive for the highest return of investment, which until recently has meant greater focus on Europe, then Asia, and is just now on a large scale expanding beyond those horizons (Schaler, 2013). However, the world is changing as more people are accessing technology whether through computers, cell phones or tablets, and markets with millions of people are opening up. It will be increasingly important for businesses to respond to this in order to take best advantage of it. Proof of this trend is simply to look at Microsoft and to see the growing list of languages offered for their products or to look at Facebook and Twitter to see the ever expanding language options. Of course, in order to offer tools in these languages such a spell checkers, there is a need to first develop linguistic resources such as a morphological grammar.

1.3 NLP and Under Resourced/Endangered Languages

Another movement that has emerged is the concern with and desire to develop linguistic resources for not only under-resourced languages with a large amount of speakers but also the push for development and support of technology in the assistance of language maintenance and

revitalization for languages that are struggling for survival. One linguistics technology resource and consulting company, Idibon is calling for NLP support for all languages. Idibon works with companies to help them understand their linguistic data and utilize NLP programs in order to accomplish their tasks (Idibon.com, 2013). This often includes encouraging and coaching companies to diversify the languages their business model can support.

Idibon is a representative member of the movement of companies that are putting energy in to these endangered languages in varying ways that also include the development of smartphone apps to support under resourced and endangered languages such as Ojibwe, Mohawk and Inuktitut. Linguists, community activists, and software developers are working together to create a plethora of different technological tools including video games to encourage use of these languages and to demonstrate their pertinence and applicability of these languages in the current world. Tusaalanga is a free iPhone application to help learn Inuktitut, which is an example of the fore mentioned movement and development of tools for endangered languages. This iPhone application was created by the Pirurvik Centre for Inuit Language that along with the development of an Inuktitut learning application provides language courses across the province of Nunavut and actively works with Microsoft for the creation of Microsoft Office and Windows tools in the Inuktitut language, such as an Inuktitut language pack (Pirurvik.ca, 2013). A different company, Ogoki Learning Systems, Inc, based out of the Sandy Bay Ojibway First Nation Reservation in Manitoba, Canada, is a smartphone application design company that has also been very much been active in supporting the language of their community, Ojibwe, with a language learning smartphone application (Ogoki Learning, 2013). Additionally, for the Mohawk language, Monica Peters, an independent software developer and a member of the Mohawk nation of the St. Regis Mohawk community similarly designed a language app for Mohawk (Talkmohawk.com, 2013). Taking it a step further, a company based out of Las Vegas, Nevada, Thornton Media in 2013 released an entire video game in the Cherokee language. Additionally projects are currently in works for a series of e-books in Cree, and products that allow speakers of these languages to interact with mainstream computer software in these native languages (News.ualberta.ca, 2013).

1.4 Endangered Languages and Natural Language Processing

What needs to be developed further in this section is the development of NLP tools for languages, specifically the discussion of tools for endangered languages, and additionally the discussion of polysynthetic languages, such as Mohawk, which pose unique challenges to computational linguists. What is more, most polysynthetic languages are either under resourced or endangered, which is an interesting development in the evolution, expansion and shrinking of languages (Baker, 1996).

Challenges facing under resourced and endangered languages are that historically most of the linguistic analysis methods and approaches have not only been tested for European languages but also primarily designed for these languages. It is important to include that to date this is not still the case as extensive research and development for language tools including some Asian languages and the diverse languages of India has occurred. Yet, it is still the case that language support for Western European languages remains very dominant. A problem with this is that the European languages may at first glance appear to be very diverse, but almost entirely come from the same language family, Indo-European and thus share many foundational structural elements. Thus a linguistic and computational approach that works sufficiently for many of the European languages, or even the more analytic languages such as Chinese and Japanese, may not work well for a language from a different language family that has a completely different structure. Linguists and developers have recognized this and are actively working on many of the non-Indo-European to still include the languages of India, languages of Africa, and recent progress has been made for Native North American languages as well (Bosch and Jones et al., 2007, p. 22). In the case of many of the languages in North America, even though there exist over nine different language families in contrast to Europe's primary one, many of the North American languages do share a common property of being highly morphologically inflected to the point that many of these languages are considered to be polysynthetic. Polysynthesis poses a challenge for computational linguists attempting to create a computational grammar of a language because each word is so highly morphologically inflected and what affects some of the inflections may come from surrounding words or other morphemes on a word, which is called a long

distance dependency. Long distance dependencies are a great challenge to NLP systems in English, for example with subordinate clauses in sentences, but long distance dependency poses an even greater challenge to NLP systems for polysynthetic languages. This is because, while in English one can write a sentence in a simpler manner to avoid the complex subordinate clause structure, however for many polysynthetic languages it is a regular and common occurrence within individual words and in sentences. Current research in NLP and in the development of computational grammars is looking closely at these polysynthetic languages, in search of a way to describe them computationally which include work done by Dyck and Kumar (2012), and Lonsdale and Masushita (n.d.). For these languages previous research of the Indo-European languages and many Asian languages do not go far enough to help in the creation of tools that can handle the unique structure of these polysynthetic languages, though some work with more synthetic languages such as Bengali and various languages of India are highly inflected to a point where progress with these languages has aided researchers of polysynthetic languages (Homola, n.d.).

1.5 Iroquois Language Family

The Iroquois language family is a family of languages native to North America that are known for their very high morpheme to word ratio, and for having a structure that by some linguistics is claimed to the most different from English (Baker, 2002). Unfortunately there are very few speakers of these languages in this family, but there is a strong resurgence in the movement to preserve and to revitalize the languages over the last twenty years by their communities. These languages pose unique challenges to computational linguists and developing NLP support and resources for these languages are vital because they can aid in the preservation and revitalization of these languages and also serve a purpose to bring NLP and language engineering to new horizons in efforts to develop approaches and methodologies to handle the highly complex morpho-syntactical structures that these languages contain, which was previously discussed (Mithun, 2005). One approach to creating NLP resources for polysynthetic languages has been using finite state morphology (*to be discussed in detail in chapter 3*), which has been successfully used to create morphological parsers and other linguistic tools for both European languages and some polysynthetic

languages and languages with long distance dependencies, which are two elements that have posed until recently mostly insurmountable to computational linguists. However, while finite state morphology has been used for a diverse range of languages including highly morphologically inflected languages, the level of morphological complexity that exists in the Iroquoian languages has not so far been successfully achieved using finite state methodology. Attempts have been made to create a network for Cayuga nouns, one of the Iroquoian languages, which proved successful but still did not result in a system that could completely contain and describe the Cayuga language (Graham, 2007). The product of this research will help the language communities and will also help linguists better understand these languages, specifically Mohawk, whose structures vary so much from the mainstream European languages in understanding the human and language development.

1.6 Finite State Morphology

In order to create a morphological analyzer for Mohawk, two-level finite state morphology will be the approach used. This is one of the most ground-breaking research efforts in NLP, which was designed in 1983 in the dissertation *Two-level model for morphological analysis* (Koskenniemi, 1983). In it Koskenniemi provided a morphological analysis of his native language Finnish. This was a great accomplishment because the Finnish language, as a member of the Finno-Ugric language family and not of the Indo-European Language family like most other European languages has a greater degree of morphological inflection and complexity, which had proved difficult for developing NLP applications. It was then in 1983 in the paper *KIMMO: a general morphological processor*, where the name 'KIMMO' after Kimmo Koskenniemi was given to his morphological analysis design (Karttunen, 1983). However, the original KIMMO needed to be used on a type of computer that was only available in research centers, and it was the development of the PC-KIMMO that allowed for linguists across the globe to attempt to create finite systems to describe other languages, which now includes Spanish, Turkish, and Quechua, among others.

It was using the PC-KIMMO parsing tool that the first Native North American language, Lushootseed, was provided with a morphological analyzer. Lonsdale and Matsushita at Brigham

Young University utilized finite state morphology and the PC-KIMMO to create an analyzer for a polysynthetic language, which was a great feat (Lonsdale and Matsushita, n.d.). It demonstrated that the finite state methodology could be used to create systems for some of the world's most complex languages. However, there are limitations for finite state and to the PC-KIMMO, which is due to the fact that a finite state can only make decisions based on the immediately preceding step before is which creates an inability to handle templatic morphology, long distance dependencies, and reduplication (Antworth, 1992).

Templatic morphology was an issue for languages such as Arabic and Hebrew, whose orthographies' only include the consonants and the spoken vowel between the consonant is determined by both its surrounding consonants and preceding and/or following vowels (Beeseley and Karttunen, 2003). Traditional finite state morphology could not handle this, nor did it have a way to represent the missing phonological aspect. This is not as much of an issue for the Iroquoian languages as vowels are included in the orthography.

However, both long distance dependencies and reduplication pose a real challenge for NLP in general, and are common constructions in the Iroquoian languages. Reduplication was already discussed, but long distance dependencies pose not only a challenge to computational linguistics, but specifically to finite state automata because these systems can only make decisions based on the preceding step, however in languages like Mohawk, an affix later in the word can be affected or determined by an affix earlier on. PC-KIMMO did not have a way to handle languages that have long distance dependencies within a word, which renders it practically impossible then to implement in the design of a parser for Mohawk or any other Iroquoian language.

While not thinking of Mohawk specifically, the solution to these issues was presented by Beesley and Karttunen (2003) with their development of the Finite State Morphology software program. Beesley and Karttunen create a compile-replace function to assist the system when handling templatic morphology and allowed for flag diacritics to help with the programming of long distance dependencies and reduplication.

The XEROX finite state software has been used to creating parsing tools for Arabic, the Bantu languages of Africa, and even an Iroquoian language: Cayuga. The program with its solution to templatic morphology allowed for the creation of a finite system that could describe Arabic. In addition, work by Bosch and Jones et al. (2007) with this software has allowed for the creation of morphological analyzers and machine-readable lexicons for the South African Bantu Languages. Most recently and most relevant is the work done by Dougal Graham (2007) successfully designed a morphological parser for nouns of the Cayuga language. Due to the level of morphological complexity of the Iroquoian languages it had previously been unknown if morphological analysis could successfully be implemented using finite state morphology. During the course of this thesis he was successful in the creation of a finite system to handle Cayuga nouns. However, due to time constraints he did not attempt verbs, which are also much more complex.

The work by Dougal Graham is important when considering attempts to create a finite system for Mohawk because these languages are in the same language family. This being said Cayuga is more closely related to Seneca and Onondaga while Mohawk shares more similarities to Oneida (Chafe and Foster, 1981). It is, however, yet to be seen if a finite state machine can not only be used for morphological analysis of a language with such high polysynthesis, such as the Iroquoian languages, and if upon completion it can run with adequate speed to prove useful in other applications such as an e-dictionary or spell checker.

Using finite state morphology is not the only approach and recently the use of Machine Learning (ML) to develop morphological parsers and other NLP applications has been successfully used for some of the major European languages and work is also ongoing for the major Asian languages as well. ML has already been successfully implemented to create parsers for some languages in India. Tamil is one of the Indian languages where linguistics used MST and MALT parsers on a corpus of 25,000 words and using machine learning were able to create a morphological parser (Kumar et al, 2010). As it is some of the most current research in this area, using ML learning to help linguists and computer scientists decipher the rules of a language and to quickly be able to

develop POS taggers based on the patterns of a language, this is an important area of growth in NLP and the approach where many researchers are placing their focus.

A concurrent trend in research and linguistic initiatives using along with ML, is the incorporation of NLP in language revitalization efforts and support for endangered or under-resourced languages. Some of these endangered languages from each of the continents have received interest and support from either major international companies like Microsoft or smaller academic organizations. An excellent example of research with Native North American languages, all of which are considered either endangered or at high risk and are under-resourced, is that of the research done by the National Research Council Canada with the development of an Inuktitut morphological analyzer, e-dictionary, and Inuktitut search engine. Much of this increased support for Inuktitut comes as the new province of Nunavut, which is primarily Inuit, passed a law that brings Inuktitut of equal value in the province as English and French. Even without support from the national government efforts are underway to support NLP of all the world's languages.

It is with the current trends in NLP and linguistics for language preservation and revitalization that I undertook the challenge to create a morphological analyzer for the Mohawk language. Using finite state morphology and the Xerox Finite State Morphology software from Beeseley and Karttunen's book *Finite State Morphology* (2003), along with three Mohawk language references, I created a lexicon for a finite state system that analyzes Mohawk formal nouns.

1.7 Guide to Other Chapters

In Chapter two there will be an explanation of the Mohawk language, both in historical significance of the Iroquois people in North America and a linguistic presentation of grammar and morphology. Chapter Three will discuss morphological analysis and provide a literature review discussing the research accomplished in this field. Following Chapter three, will come the presentation and design of my experiment to create a finite state morphological parser for Mohawk formal nouns, including how Mohawk became the language of focus and what resources I was able to gather and how this affected the trajectory of the project. In Chapter five there is a discussion of the

experiment and the results, and Chapter 6 is where I will present my conclusions and what suggestions for future study of the Mohawk language. Following these chapters will follow the references used for the project and in the appendix is where the code I created for my parser will be provided.

2. Status of the Mohawk Language

2.1 Language Families and North America

Unlike Europe where most of the languages have a common ancestor, Proto Indo-European, and therefore are members of the same language family, in the North American continent it has been theorized that at least nine different language families exist, to include: Algic, Iroquoian, Muskogean, Siouan, Uto-Aztecan, Athabaskan, Salishan, Eskimo-Aleut, Mayan, among others. *See figure 2.1*

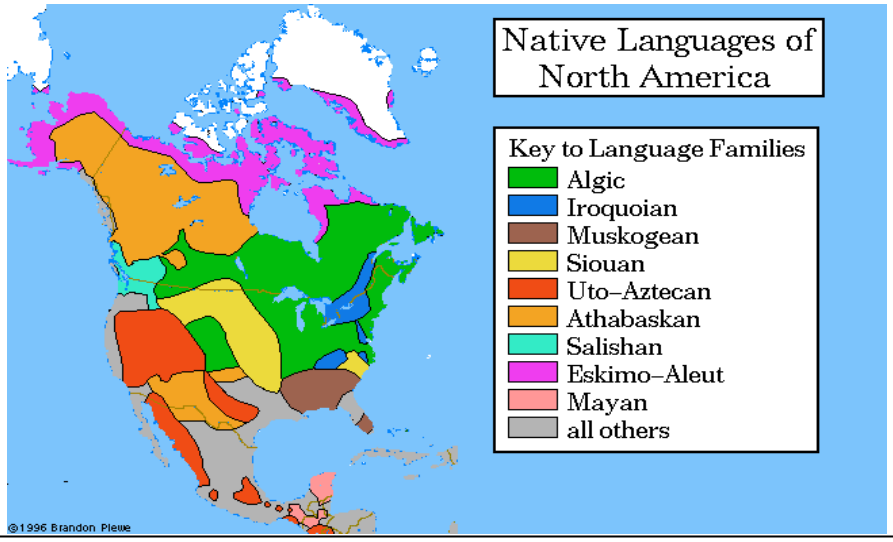

Figure. 2.1 Map of the North American Continent showing the proposed distinct language families that exist and approximately where they are spoken (Aaanativearts.com, 2013).

Perhaps the two most researched and studied language families in North America are the Algic family, which includes Cree and its varying dialects, and the Athabaskan family, which includes Navajo. Additionally, due to the recent creation of Nunavut as a new province of Canada and the status of Inuktitut as one of the official languages of the province on equal par with French and English, this language has received the attention of linguists and government for research and language resources development.

However, while North America has rich linguistic diversity, many of the languages spoken on this continent are struggling to survive, having been subjected to up to four centuries of prejudice, discrimination and eradication often institutionalized and supported by government legislation by the

colonial governments and the independent governments of Canada and the United States. According to Statistics Canada only Inuktitut with 34,000 speakers, Cree with 16,000, and Ojibwa with 30,000 speakers are expected to survive until the next century (Www12.statcan.gc.ca, 2013). Meanwhile in the United States, the Navajo language, which has approximately 150,000 speakers and is the language with the most speakers of all the Native North American languages, is still considered a threatened language according using the United Nations scale for language health (Technaverbascripta.wordpress.com, 2013). As for languages with a few thousand speakers, the predicted future is bleaker still.

Another important factor that affects the current language efforts is that many of the native North American languages had no written record until the middle of the 20th century when linguists began collecting linguistic data (Dyck, 2009). What this means is that linguists have limited information about many of the languages of the North American continent, and that creating resources for these languages has been challenging due to the lack of written records, a complete grammar, and a dwindling population of speakers with native level or advanced level knowledge of their culture's language.

The linguistic diversity of North America is an opportunity for linguists to research languages with patterns and parameters that differ highly from the Indo-European languages and other language families in the world. There are not many languages that are considered to be polysynthetic, which is a linguistic typological classification for languages with a very high morpheme to word ratio, yet many of the languages in North America are considered to be polysynthetic. To put this in perspective, languages like Chinese and Japanese are considered analytic languages because each word has little to no morphological inflection. In this case, the verb might not conjugate and so the subject and tense are not expressed by morphological changes within the verb word, but in separate words as markers. In comparison then English has greater morphological inflection than Chinese and Japanese, but in many ways has analytic properties as well since to make the future tense we do not change the verb but we add a construction with another word, "I *will* run." Then, the romance languages have more morphological inflection than English, but less than German and Russian which

are considered to be synthetic languages. Synthetic languages have rich morphological inflection, and at this stage someone looking at a word in these languages searches for the root of the word, which will then be surrounded by varying affixes. Russian for example has far fewer prepositions than English, but this is because within the word in a sentence its role has been defined by some marker. This makes the necessity for prepositions not obligatory because dative, instrumental, or genitive cases are reflected by changes that occur in that word. Languages such as Inuktitut, Cree, Ojibwe, other Algonquian languages, and the Iroquoian language family are all considered to be polysynthetic languages. Some of the characteristics of this language typology include noun incorporation, and long distance dependency, and are often described to people as the type of language where one word in these languages often translates to an entire sentence in English. For this reason polysynthetic languages are often simply described as languages with sentence words. Additionally, fewer parts of speech are necessarily such a prepositions, adverbs and even adjectives since semantically they are represented as a morpheme within a word.

2.2 Iroquoian Language Family

The Iroquoian language family, all of whose languages are considered polysynthetic, is one of at least nine distinct language families in North America. The family's two main branches are Northern and Southern Iroquoian. The Northern Iroquoian languages are spoken across upstate New York and Michigan in the United States, and in parts of Quebec and Ontario, Canada. However, the most well-known language and the language with the most speakers and resources is Cherokee. Cherokee is a Southern Iroquoian language, in fact the only Southern Iroquoian language, and it is spoken in regions of North Carolina and Oklahoma. This language has significantly more fame from the historical Cherokee Trail of Tears, a forced displacement and relocation that killed many and forced others to suffer, which was the result of the 1830 Indian Removal Act that sent the Cherokee and other nations from their traditional territories to lands farther west. The primary reason why Cherokee is spoken in Oklahoma is a result of the displacement of the Cherokee nation at that time.

While the Southern branch consists of only Cherokee, the Northern Iroquoian branch includes the primary languages of Mohawk, Oneida, Onondaga, Cayuga, and Seneca. *See figure 2.2*

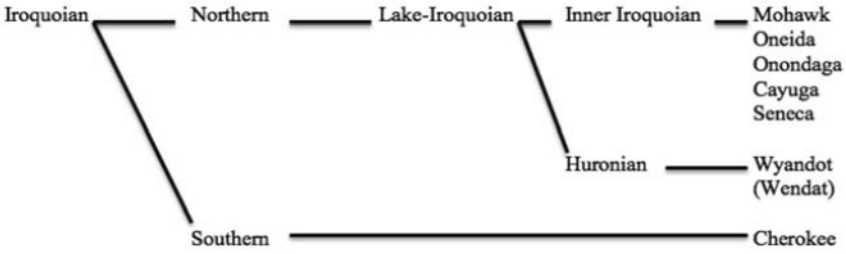

Figure 2.2 Iroquois Language Family and branching structure (Dyck and Kumar, 2012)

Additionally, the Tuscarora language is considered to be an Iroquoian language but is proposed to have broken off from Proto Northern Iroquoian before the evolution to Lake-Iroquoian and was not included in figure 2.2 (Chafe and Foster, 1981). All of the Northern Iroquoian languages are very similar to one another and in 1952 Hickerson, Turner, and Hickerson evaluated the mutual intelligibility of the Northern Iroquoian languages. They found the mutual intelligibility between Mohawk and Oneida to be upwards of 80 percent or more, and between Seneca and Cayuga to be between 75-80 percent or more (Hickerson, Turner, and Hickerson, 1952). Onondaga stood relatively isolated and across all of the five languages as it was between calculated to be around 35 percent. This supports the work of Chafe and MFoster (1981) where they proposed a series of divergences and convergences between the Iroquoian languages and claimed that Seneca and Cayuga share the most recent common predecessor, and that Mohawk and Oneida were once one language. Additionally, due to movement among the tribes and boundaries, at times the languages experienced divergences in their evolution and then through periods of re-contact almost converged into one language. Robbie Jimerson, a member of the Seneca tribe recalls his grandfather telling him about working in New York City building the skyscrapers and that he would converse with men from the Mohawk nation (Jimerson, 2013). While the Mohawk speakers could not understand the grandfather's Seneca, he could understand the Mohawk of his colleagues. Indeed Wallace Chafe, a linguist who first worked with the Seneca tribe in the 1960s and wrote the first morphological and syntactical literature about Seneca theorizes that for much of the Iroquois confederacy, the different tribes spoke a common

Iroquoian language. He argued that Mohawk is the most conservative and has evolved the least from that common Iroquois language, whereas the farther west one goes in upstate New York, the more changes and innovations occurred in the Iroquoian language. *See figure 2.3 for a map of New York State.* When analyzing the words in Seneca, in many cases the root of the word has disappeared altogether, but its heritage root can be estimated from 17th century Mohawk language documents. Additionally, the Seneca tribe, considered the Western Gate Keepers, received linguistic influence from their neighbors the Huron, and from languages in the Algonquian language family.

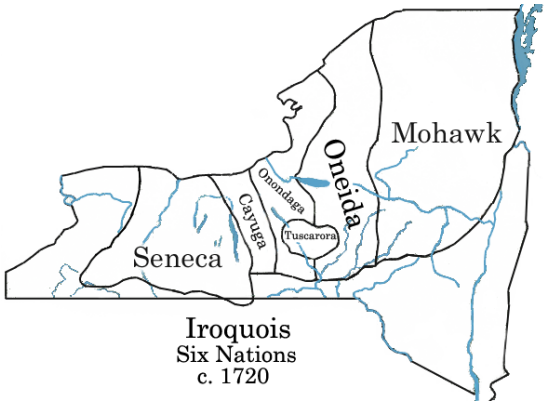

Figure 2.3 Map of New York State and Iroquois Six Nations territories (Iroquoismuseum.org, 2013)

2.3 The Historical and Linguistics Relevance of the Mohawk Language

The Mohawk language and the Mohawk nation as a great whole has been a player in North American History both before and after the arrival of the European settlers. It is theorized that the Mohawk were some of the first indigenous peoples to have contact with the Europeans in 1534 near what is today Quebec City, Canada. Jacques Cartier called these people the Laurentians and it is from these people that Canada gets its name: *Kaná:ta'* which is still a noun in contemporary Mohawk that means 'settlement' or 'town'(Mithun, 1996). The Mohawk tribe belongs to the Iroquois Confederacy that was historically one of the most feared and powerful Leagues of Nations in North America and in the 16th and 17th centuries was responsible for pushing many tribes out of the northeast, such as the Sioux who were originally from what today is the state of Ohio but who are now famous for being a fierce plains tribe that fought ardently against the United States' western expansion (Charles River, 2012).

The Iroquois Confederacy, or League of Six nations, calls themselves the *Haudenosaunee* which means, "People of the Longhouse" based on the type housing they constructed and in which they lived. It is also based on the layout of the six nations in the region of what is today Upstate New York. *Please refer to figure 2.3 for reference to the geographical organization of the Iroquois Confederacy*. From East to West were the Mohawk, Oneida, Onondaga, Cayuga, and the Seneca, who were the western gate keepers. The term 'Mohawk' is actually a name given to the tribe by their neighboring tribes and enemies. It is thought to be a Narragansett word, which was an Algonquian language that is now extinct, and can be roughly translated to "people who eat living things," which has been interpreted as *cannibals.* The Mohawk refer to themselves as *Kanien'kehá:ka,* which means 'People of the Flint'. The Mohawk also refer to their language as *kanien'kéha* (Mithun, 1996).

The Iroquois Confederacy was not only a fierce nation of tribes but also influenced the American Constitution, which incorporated elements from their Great Law of Peace. Through the colonial struggles and wars with the French and British, the Mohawk and the Iroquois Confederacy often offered aid and support to colonialists in the regions of New York. However, during the American revolutionary war the League of Six nations was divided in their allegiance to either the British or the colonists. Officially the Mohawk tribe supported the British Army and thus when the United States gained its independence George Washington expelled the Mohawk tribe from their traditional lands. The Mohawk and many other Iroquois people ended up settling in Canada but were divided from that point forward. The ramifications of this point in history led to the separation of the Mohawk community spread across upstate New York, Quebec, and Ontario that has resulted in the insurgence of different dialects among this small linguistic community.

Of additional and linguistic importance, the Mohawk language perhaps because of its early contact with European settlers, was one of the first North American languages to be recorded. 17th century priests and missionaries devised dictionaries and even translated the Bible into Mohawk in order to convert its members to Christianity (Charles River, 2012). In 1644 Johannes Megapolensis included a section, "A short account of the Mohawk Indians" in the publication *Early Vocabularies of Mohawk*, which was essentially one of the first written records of Mohawk ever documented

(Mayndertsz van den Bogaert et al, 1999). In addition to his list of Mohawk words, he even attempts to describe the Mohawk language with his limited experience and only having ever been exposed to Indo-European languages. In this early work, he already understood and noticed that the words would change depending on the sentence and an object's role in the sentence.

2.4 Current Situation of the Mohawk Community and its Language 2.4

The Mohawk community as of 1999, is around 30,000 people, who are living in communities scattered across New York State, Quebec and Ontario, Canada (Ethnologue, 2013). Among these communities the most important are Kahnawà:ke in Caughnawaga, Quebec, Kanehsatà:ke in Oka, Quebec, Ahkwesáshne in St. Regis, New York, and Tyendinaga and the Six Nations Reserve in Ontario. *See Figure 2.4*. As a result of the separation of these communities it has resulted in the development of three major dialectal variations: Eastern (Kahnawà:ke and Kanehsatà:ke), Central (Ahkwesáshne), and Western (Tyendinaga and the Six Nations Reserve).

Figure 2.4 Iroquois territories in modern day. The Mohawk territories are highlighted blue (Iroquoismuseum.org, 2013).

Of the 30,000 residents it is estimated that fewer than 10,000 people speak Mohawk and only about 3,500 fluent speakers exist (Ethnologue, 2013). However, while this language is highly

threatened and not one expected to survive in to the 22nd century as per Statistics Canada, of the Northern Iroquoian languages it has the greatest number of speakers (Statistics Canada, 2013). After Mohawk, Cayuga has 250 speakers and is deemed a moribund language by Ethnologue and the United Nations language vitality scale. Oneida has 192 speakers and is classified as shifting. Seneca has 100 speakers and is also classified as shifting while Onondaga has only 52 speakers. Tuscarora has only 9 speakers left and is classified as nearly extinct. During an interview with Mary Kay Olan of the Mohawk Nation, she shared that in 2000 *Haudenosaunee* gathered to discuss the situation of the language loss in each of their communities (2013). 10 years later when they gathered again, only the Mohawk community had successfully implemented programs and efforts to stop and reverse some of the language loss. Efforts in the second half of the 20th century have included the creation of the Kahnawake Education center which founded the Kahnawake Survival School in 1980 and in 1979 the Akwesasne Freedom School which in 1985 implemented a full Mohawk immersion program for the community's children. Additionally, in 1999 the Onkwawenna Kentyohkwa (Our Language Society) was founded as a community-based organization that has taught Kanyen'keha (the Mohawk Language) to adults on the Six Nations Grand River Territory in Ontario, Canada. The two year immersion program aims at teaching young adults and members that will go on to teach and utilize their language skills within their communities. Mainstream secondary schools and universities in Ontario and Quebec are also currently developing and offering introductory Mohawk and Cayuga language classes.

Linguistic and language learning resources have also been developed and published from each of the three major reservations that each proposes their own dialect. *Mohawk A Teaching Grammar* written by Nora Deering and Helga Harries-Delisle (Deering and Harries-Delisle 1976) was first printed in 1976 and comes from the Kahnawake reservation in Quebec. The most recent addition in 2007 does take into account the differences in pronunciation and spelling dialectal variation among the Quebec, Ontario, and New York speakers. David Maracle, who is based in Ontario, has written *Let's Speak Mohawk* (1993) and *One Thousand Useful Mohawk Words* (1992). However, the first modern grammar ever documented, created for linguistic purposes and which was more than just a list

of words or a direct translation of the Bible, for instance, was written by Nancy Bonvillain in 1974 with her doctorate thesis *A Grammar of Akwesasne Mohawk*. In her dissertation she was able to address and break down many syntactical and phonetically elements of the variation spoken at the Akwesasne/St. Regis Mohawk reservation. It was this source that I was planning to base the majority of my morphological analysis on; however her thesis, while excellent at classifying many nouns and verbs, was also an oversimplification, and lacked consistency among the other language resources, both of which were more contemporary and written by fluent speakers of Mohawk. Ultimately, I decided to attempt to assess the classifications using all of the major resources, using what overlapped between *Kanyen'keha Tewatati (Let's Speak Mohawk), Mohawk A Teaching Grammar,* and *A Grammar of Akwesasne Mohawk*.

What quickly became apparent, when examining the morphosyntax of Mohawk was that creating a morphological analyzer for the entire Mohawk language would be too complicated for the time and scope of this dissertation. Mohawk verbs with their multiple prefixes and affixes, long distance dependencies, and reduplications were just too complicated to attempt in this go. Even Mohawk nouns can be incredibly complex. However, the Mohawk noun presented an interesting NLP challenge and offered an opportunity to learn about morphological analysis and to become involved with the language revitalization and language tool developing trend.

3. Mohawk Grammar

The Mohawk language is classified as a polysynthetic language due to its high morpheme to word ratio. In many cases the translation of one word in Mohawk translates into a full sentence in the English language. This brings up the question of how to define a word in Mohawk and among the Iroquoian languages. Carrie Dyck (2009) addresses this very issue. Like Cayuga, all Mohawk words must contain only one stressed vowel. Additionally, when the morphemes are separated from one another and presented to a speaker he or she is unable to identify their meaning. If they were separate words, then there should be a meaning independently of the context.

In the Mohawk language there exist three parts of speech: particles, nouns, and verbs. Prepositions, adjectives, and other connectives are non-existent in Mohawk or the other Northern Iroquoian languages. How to express what prepositions and adjectives are used for semantically must be constructed in a different way. However, it is interesting to note that Cherokee, a southern Iroquoian language does have adjectives and thus demonstrates that the adjective disappeared from the Northern Iroquoian languages after they split from Cherokee.

In Mohawk there are only 12 letters in the alphabet: *t, k, s, n, r, w, ‘ , h, i, e, a, o* . In this alphabet the symbol ‘ stands for a glottal stop. Additionally, when looking at varying print resources for Mohawk in the western dialect a *y* exists in place of *i*. There are also diacritic marks for stress and tone `, ´ and vowel length : (Mithun, 1996). No bilabials exist, such as *p* or *b*, which is highly unusual in comparison to most of the world’s languages. Additionally, the phonemes *on* and *en* are nasalized. A regional variance is the pronunciation of *r*, which sometimes is pronounced like ‘l’. This occurs within the Akwesasne region, and in the Six Nations Reserve, while in Kahnawake the pronunciation is closer to an ‘r’.

Particles in Mohawk are words that have none to almost no internal morphological structure and are used in Mohawk as numbers, demonstratives, adverbials, conjunctions, exclamations, and more. Particles are usually much shorter and contain only a few syllables. *See Table 2.1 for examples.*

Mohawk Particle	English Translation
Oh	What
Tanon	And
Á:re'	Again

Table 3.1 Mohawk particles

3.1 Verbs

The most important part of speech in the Iroquoian languages is the verb. These languages are considered verb heavy and in many cases only the verb is needed to express an entire sentence semantically, containing information about the subject, object, location, tense, and more. The verb root is therefore incredibly important, but even native speakers often cannot identify the root of the verb though they understand the word without difficulty. If a native speaker cannot easily identify the verb root and detail the specific meaning of each morpheme, it demonstrates the challenges for a language learner when they come across a verb they do not understand and the difficulty facing a student of Mohawk as a second language.

A verb root can be a few syllables long, but can also be as short as -t- 'stand' or –k- 'eat (Mithun, 1996). In addition to the verb root, a verb in Mohawk must contain the following semantic components that specify the action or state, the subject and the object, and the modifications of the action or state such as mode, aspect, time, space, number, and more. Thus the most basic general structure of the morphemes required for a possible verb in Mohawk is:

Verbal Prefix + Pronominal Prefix + Verb Base + Verbal Suffix

The verbal prefixes contain the modal and non-modal prefixes, whereas the pronominal prefixes specify the subject, object, and possibly the co-occurrence of a particular subject and/or object. The verb base includes not only the verb root but also any reflexive components and incorporated noun roots. The verbal suffix contains specific verb root suffixes, aspect and attributive suffixes. The

Verbal Prefix and Verbal suffix classes are not obligatory in certain instances and depend on the classification and type of the verb root, whether it is an action or state, and if it has a beneficiary or not.

Since there are no adjectives in Mohawk, how to express the meaning of an adjective without having this part of speech requires Mohawk to find other ways to express this. For example, in Mohawk 'the dog is red', would literally be translated to 'the dog reds'. Indeed, many adjectives in English are expressed in Mohawk as a suffix within the noun.

3.2 Nouns

The noun in Mohawk is the second most important part of speech, and is less morphologically inflected than the verb. The basic form of the noun is:

Nominal Prefix + Noun Stem + Nominal Suffix

The nominal prefix contains prefixes that mark the word as a noun and sometimes denote and/or modify the noun in terms of person and number. Additionally, the nominal prefix can mark the relationship to specific verbs and to other nouns. The noun stem in turn can contain the noun root and even multiple noun roots. The nominal suffix class can mark the word as a noun, and modify the noun in terms of number, person, attributes, location, and also tie the noun to a specific relationship with a verb. The suffixes which denote location are the morphosyntactical answer to the lack of prepositions in the Mohawk language. The attributive suffixes can be used in a way to compensate for the lack of adjectives and are used in addition with verbs to modify nouns and situations. Nancy Bonvillain (1973) explains in her research that she found the nominal suffix not to be obligatory, but she also included a morpheme of 'a' in the noun stem that can sometimes appear. It would seem that this sometimes is included between the noun root and the nominal suffixes, but it also exists on its own at the end of a noun word. Other books of Mohawk consider this to be an additional and common nominal suffix and not an intermediary morpheme. Keep in mind that not just one suffix can appear, but multiple suffixes. However, from the three resources used it was not possible to determine the structural hierarchy of the suffixes and the correct order in which they are attached to the noun root.

Within the part of speech of nouns there are two major categories: formal and functional nouns. The majority of nouns are formal and describe the bulk of concepts from objects, places, and ideas, and are fairly short in length and follow the three part structure proposed above. Functional nouns are more complicated and are often used to describe professions and many newer words that have been created to describe the contemporary world. Many of the functional nouns are actually structurally considered verbs, but are used within the sentence and through context act as a noun. *For example see table 3.2.*

Mohawk	English
Ratorats	He hunts
Riyenteri ne **ratorats**	I know **the hunter**

Table 3.2 example of Mohawk verb also being a functional noun depending on context (Maracle, 1993).

In this case *ratorats* is a verb in the present tense above, and below the same word is used as a noun to describe 'the hunter'. For the purposes of this dissertation I will not be attempting to include the functional noun in my parsing program, nor will I attempt to describe the nuances of the functional noun here. Most functional nouns follow the structure laid out in the verb category and can be very long in length.

3.3 Formal Nouns

Formal nouns are typically 3 or 4 syllables. An interesting point to note is that usually one word in Mohawk has multiple possible translations in English depending on the context. In David Maracle's book *Kanyen'keha Tewatati* (1993) he explains that formal nouns can be broken down in to two more categories Category I are nouns which describe an object or place which can be influenced or changed by humans. Thus all of this vocabulary includes possessions such as, artifacts, tools, and manufactured goods. Category II includes nouns which cannot be changed by humans. This would include words for the natural environment, physical and emotional attributes, body parts, and animal names. An example of a functional noun is the word for "stove polish":

Yontenonhsa'tariha'tahkwatsherahon'tsihstatsherahstara'the'tahkwa
The stuff that makes shiny that one puts on thin that is used to heat the house

Table 3.3 example of sentence word in Mohawk (Maracle, 1993).

Category I nouns usually take nominal prefixes of *ka, a, aw, e/en,* and *o*. In speech, often *e/en* prefixes are preceded by *aw*. While David Maracle creates the 2 categories, Nancy Bonvillain's dissertation does not coincide with the classification scheme he used. She only includes the prefixes ka- and a-, whereas Maracle proposes 3 to include ka-, e/en-, and o-. For the purposes of my morphological analyzer and to attempt to include parsing for more words I follow Maracle's schema and include aw- as a fourth prefix and the null prefix as a fifth possibility.

Ka-	A-	E-	O-
Kanonhsa	Ahta`	Erhar	Onenshte
Kanakta	Athere	Ehsa	Owira
Katshe	Atoken	Eryahsa **Aw**eryahsa	

Table 3.4 Examples of four nominal prefixes for formal nouns.

Bonvillain also addresses a set of nouns that do not follow the formal noun structure, and includes them in a class of their own. Some of these words include, *kitkit (chicken), ehlalh (dog), tako:s (cat), and eshah (black ash tree)* (Bonvillain, 1973). What is interesting is that these nouns are classified under Maracle's organization scheme under a mixture of category I and II nouns. For instance the words that begin with 'e' might get classified as a category I noun, yet they do not follow the typical structure. Bonvillain chooses to propose that many animals are indeed words that do not follow the formal structure of nouns, and that perhaps some of the names are onomatopoeic (Bonvillain, 1973). Indeed in *Mohawk A Teaching Grammar*, many animals, insects, occupations, and kinship terms are not classified as specifically formal nor functional nouns because they do not necessarily and uniformly follow that schematic either (Deering and Harries-Delisle, 2007). This group of nouns that is unique in its properties and how morphemes are added to it I will call defective nouns in my morphological analyzer.

Within the noun stem, in addition to the noun root Bonvillain includes what she proposes may be an additional and optional morpheme joiner, which she says is usually /a/ (1973). Yet, her dissertation is the only source that even discusses this morpheme joiner, but perhaps it has to do with certain orthography and pronunciation rules that cause this optional phoneme to appear in words depending on the last letter or syllable of the noun root and the beginning letter or syllable of the nominal suffix that follows it. The basic nominal suffixes as per Bonvillain are /‘/, /a’/, or /u’/ but according to Maracle there is only /a/ or /e/ as the most basic nominal suffixes. These most basic suffixes create a verbal aspect along with assuring the word’s classification as a noun so that one could translate the noun with two meanings. *See Table 3.5*

Kahyatonhsera	(it) book (is)
	(it’s a) book

Table 3.5 translation of Mohawk word ‘book’ (Maracle, 1993).

The most common suffixes are for pluralisation, location (that express on, at, in, etc), attributive (size, quality, and newness among others). Thus to find the noun root of a word one needs to be able to identify the nominal prefixes and suffixes and what is left is the noun root.

Kanata -> (ka) nata -> nat (a) -> -nat-

What must also be considered when combining and parsing words is that certain spelling rules must be followed. An example is that when /à/ appears before a *k, t,* or *s* and that syllable is not receiving the stress, then the /à/ will appear as /a’/.

Àthere -> àther(e) -> a’ther-

3.4 Possession

Demonstrating ownership in Mohawk involves the use of a set of prefixes that depend on whether the noun begins with /a/ or a consonant. It is important to understand that these prefixes are never used with animate nouns. This is because the concept of a noun representing a living creature or being is such that it cannot be possessed by another living creature. Additionally, this is where the issue of alienable vs. inalienable nouns presents itself. Inalienable nouns are nouns such as body parts

which can never be separated or "alienated" from an individual. They take their own set of possessive prefixes, but will not be discussed here because for my system I did not include them due to the complexity of their structure, which incorporates the rules of both formal and functional nouns. See Table 3.6 for the possessive prefixes for noun roots that begin with either /a/ or a consonant.

Mohawk prefixes for words that begin with an /a/	English	Mohawk prefixes for words that begin with a consonant
Akw-	My	Ake-
s-	Your	Sa-
rao-	His	Rao-
akao-	Her	Akao-
ao-	It's	Ao-
onky-	our (two of us)	Onkeni-
tsy-	your (two of you)	Seni-
onkw-	Our	Onkwa-
sew-	your (all of you)	Sewa-
raon-	their (male)	Raoti-
aon-	their (female)	Aoti-

Table 3.6 Possessive prefixes for Mohawk formal nouns.

3.5 Pluralisation

Pluralisation occurs on nouns normally by using a suffix, and depending on the number and the specific object sometimes a modifier is placed before the noun in addition to the pluralisation suffix. However, while Category I and Category II formal nouns have loosely defined laws, another scheme is used when attempting to pluralize nouns based on whether the nouns are animate or inanimate. The plural suffix for animate nouns is –okon and for inanimate nouns is –okonha (Maracle, 1993). Additionally it is possible to see –okonha written as okon'a in some dialects.

Some animate nouns require a change in the pronominal prefix, however, these are nouns for people, and kinship terms that I will not be analyzing and therefore will not explain here. It is important to mention though that this change does occur and needs to be addressed for development of a complete morphological parser. An example would be the word for 'man', ronkwe, which when pluralized needs the prefix *ro**no**nkwe*, 'men'. An example of an animate noun that not only needs a prefix to pluralize it but both is 'boy', raksa'a, which to make it 'boys' is *ra**ti**ksa'**okonha***. In this case it takes the expected plural suffix for animate nouns but also needs the prefix ti- and it is important to note that this prefix precedes the noun root –ksa'a- but follows the ra- nominal prefix.

What complicates the task of pluralisation further is that some dialects use a different suffix –hsonha that works for both animate and inanimate nouns. Addtionally, it can also happen that the suffixes –okon/okonha and –hsonha can be used for the same noun to denote subtle differences in the type of pluralisation.

Kahyatonhsera**'okon**	books – refers to a grouping of books that may be the same type of size, or topic
Kahyatonhsera**hsonha**	Books- refers to a group of books that are different in type, size, or topic

Table 3.7 comparison of meaning for pluralisation suffixes (Maracle, 1993).

Another plural suffix –hshon is used on nouns that already have a locative suffix and succeeds the locative prefix.

3.6 Locative Suffixes

The locative suffixes are attached to a noun to express what a preposition would be used for in English. This covers *in, to, at, on, within, under, near, beside,* or *next to*. The suffix -àke is translated as *on* but can also be interpreted as *in, to, or at* depending on the context.

Kahyatonhser**àke** – on the books

An alternative to –àke is –hne, and the nouns that take in rather than – àke must be learned as exceptions.

Ennitskwahra**hne** – on the chair

The suffix -akon is the suffix that is most commonly translated at in, within, but can also be interpreted as within.

Kahyatonhser**akon** – in the book

The suffix -òkon is interpreted at *under* and can also be translated at *beneath* or *underneath*.

Kahyatonhser**òkon** –under the book

The locative -akta denotes an object as being beside it. It can also be interpreted as near or next to.

Kahyatonhser**akta** – near the book, beside the book

For most nouns simply deleting the nominal suffix –a and replacing it with the locative suffix is sufficient, however some nouns require that before a locative suffix is added that another morpheme connector is placed in between the noun root and the locative suffix –atsher- (Maracle, 1993).

Ennitskwara-> ennitskwahr**atsher**òkon

3.7 Attributive Suffixes

The positive attributive suffix is –iyo and can be translated as *nice, good, suitable, useful* among others. The negative suffix is –aksen and can be interpreted as *bad, poor, useless, awkward, unacceptable, unsuitable, or inappropriate* depending on the context and the noun. The next pair is for oldness and newness, -akayon and –ase respectively. They can be rendered as ancient, archaic, and also recent or fresh. Size however, is more complicated. The suffix –owanen can be used to denote something large, but both a prefix and suffix is needed to mark it as small in size, ni-a'a. (Maracle, 1993). Additionally, the suffix –ko:wa can be used when an object is large for its kind, and not just large in its nature (Maracle, 1993).

3.8 Other affixes

There is also a suffix marking if something is genuine, real or true. –on:we. Numbers and counting objects also poses a complex situation for Mohawk, since some ways to denote number can be added as a prefix, a prefix and suffix, or an additional and separate word plus a special number suffix. There is also a specific way to add and incorporate multiple affixes to the noun root, but it seems that the system for this has not been precisely described nor have rules been provided to assist in understanding the greater system. However while this information was not available, like the general structure of both verbs and nouns, the morpheme order is rigid and invariant (Mithun, 1999), which makes it likely that there would be a strict hierarchy and order for additional affixes.

4. Research in Morphological Analysis

4.1 Morphological Analysis

Morphological analysis of a language is needed for natural language processing (NLP) of that language. Programs and software that utilize NLP are in many facets of technology from iPhone auto correction, to search engine machines, speech recognition programs, spell checkers, and electronic dictionaries. Morphological analysis can be accomplished using various formalisms of grammar and depending on the typology of some languages morphological analysis may be more or less straightforward though human language has proven to be anything but straightforward even in a language's simplest form and for the most analytical languages. Analytical languages and languages that use morphological inflection less to denote meaning and assign roles in sentences do provide a less complex problem for computational linguists. Synthetic languages and polysynthetic languages with very high morphological inflection provide greater challenges since word order, and changes within each word must be analyzed to determine meaning.

The languages with the most NLP support are western languages, and the languages of the Indo-European family. This includes English, Spanish, French, among others. However, while German and Russian are more highly inflected, they have received attention from linguists with successful efforts to create computational grammars. However, while German and Russian provide a greater challenge, due to the typological differences between most polysynthetic languages and the Indo-European languages it has been a challenge to utilize the most mainstream theories of formal grammar to describe them, let alone utilize theories and practices to create analyzers and computational grammars. Additionally, the focus on developing NLP tools for languages began and is still directed mainly by businesses and Return on Investment projections by the world's leading companies. While there are polysynthetic languages with many speakers, the situation across the globe is such that many polysynthetic languages are under resourced, lacking in written documentation and standardization, perhaps having never been recorded or documented, and linguists know very little about these languages.

To describe and perform linguistic analysis on most polysynthetic languages many researchers have utilized Lexical Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001). This formal grammar considers languages to have multiple layers and dimensions to its structure, and has been useful in analyzing languages with free word order and high morphological inflection. There is both an f-structure and c-structure, which represent the grammatical function and the syntactical constituents respectively. It is using LFG that Petr Homola (2012) discusses morphological analysis of polysynthetic languages in his paper *Machine Translation Toolchain for Polysynthetic Languages.* He addresses how to use LFG to create a morphological parser for Aymara, Guaraní, Quechua, Abkhaz and Georgian, none of which are Indo-European languages and express varying levels of synthesis, polysynthesis, and other unique features. His focus in *Parsing a Polysynthetic Language* (2011) is about Aymara, an Amerindian language spoken in Bolivia, Chile, and Peru. There are approximately 2 million speakers and it is a major linguistic community in the South Cone of the Americas. Much like Mohawk, Aymara words in translation often encompass an entire sentence in English. Homola uses the example, *alanxarusksmawa*, which means "I am preparing myself to go and buy it for you" (Homola, 2011). In the development of a parser for Aymara he must overcome features in Aymara that have traditionally prevented linguists from creating NLP grammars for these languages such as Free Word Order, polypersonal agreement, and vowel elision. To accomplish this he uses Lexical Mapping Theory (Bresnan, 2001). This is an incredibly cumbersome and complicated approach involving many layers and structures of a sentence in order to accomplish the morphological analysis. He includes analysis in c-structure, f-structure, i-structure, and a-structure, which covers word order, dependencies and co-references, topic-focus articulation and valence respectively (Homola, 2011).

Antony and Soman in their paper *Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey* (2012) do an excellent job discussing the challenges that many of the Indian languages face in terms of development of NLP systems. Many of the languages are agglutinative and are highly morphologically inflected which as discussed above makes morphological analysis more difficult, yet it is one of the intermediary and crucial steps toward the

development of bilingual and multilingual Machine Translation (MT) systems. In their paper they outline the varying approaches to developing morphological analyzers. They discuss rule-based, corpus-based, and algorithmic.

The rule-based approach is what Petr Homola (2011) utilized with the LFG formalism, and he is not alone in using a similar rule-based approach to tackle the complicated morphology of polysynthetic languages. Naira Khan and Mumit Khan (2006) completed a project using Head-driven Phrase Structure Grammar (HPSG) to develop a computational grammar for Bengali. They implemented the Linguistic Grammars Online (LinGO) initiative at Stanford University which has open-source HPSG computational resources and the Linguistic Knowledge Based Grammar Engineering Platform (Khan and Khan, 2006). Bengali is considered a fusional language and not polysynthetic, but is not one of the major European nor the major Asian languages that have been previously and currently researched to a high extent. Like Aymara and the other polysynthetic languages Petr Homola addresses using LFG, Bengali has also been analyzed using LFG.

At the Kitami Institute of Technology and in collaboration with Hokkai Gokuan University in Japan, researchers work with the Ainu language of Japan (Ptaszynski and Kazuki, 2013). The project included developing a part of speech tagger which then led to a tokenizer and development of morphological analysis capabilities, which they hope can be used for future machine translation programs. The first product they developed was called POST-AL (*Part-of-Speech Tagger – Ainu Language)*, which they created using a dictionary that had already had a well-structured part-of-speech classification system (Ptaszynski and Kazuki, 2013). This aided them in creating a tokenizer and for use in development of a morphological analyzer. Ainu is an endangered language and also a polysynthetic language, in contrast to English, which is more analytical and has relatively little morphological inflection. Words in Ainu can be comprised of many affixes that surround the root of the word. What helped these researchers was having a resource of Ainu grammar and classification that was well developed such as the dictionary, which was able to be converted to XML and used in the process for the development of an analyzer. In many cases, endangered languages have very little resources or records, and a grammar has not been well-classified or described in any standard way,

which is inhibitive to the development of a computational grammar and tools for linguistic analysis. This project with Ainu demonstrates what is possible and can be achieved for polysynthetic languages.

Another approach to morphological analysis is Finite State Two-Level Morphology. This approach was first developed by Kimmo Koskenniemi, a Finnish computer scientist in 1983. His approach treats the morphological analysis as a system composed of states, transitions and actions. The Finite State Machine/Finite State Network has one start state and one or more final states. What is key in their design is that the Transition between states is only possible if the required input is recognized as a part of the system. What is unique about these systems that differs from other rule-based approaches is that rules are symbol-to-symbol constraints that are applied in parallel, not sequentially like rewrite rules in other rule-based systems. The two-level morphology then allows the system to interpret a word as its surface level and its lexical level. *See figure 4.1*

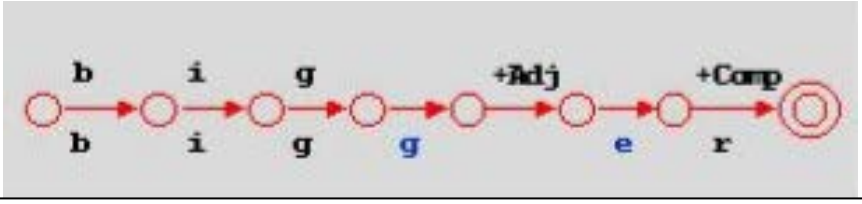

Figure 4.1 Example of finite state two-level morphology on a word (Megerdoomian, 2013).

The system addresses each word independently. Each string is one word, and a language consists of a set of strings. Each word should ideally only have one specific path through the network. Since there is a surface and lexical form, the design of the program is to work bi-directionally. This creates many advantages for bilingual dictionary and machine translation work. The two-level morphology can allow the inclusion of the multiple morphemes, where the absence or presence of one leads to an altered meaning of the word. The ultimate goal of a finite state morphological parser is that it accepts all valid words in the language and rejects invalid words by not returning any output. It does this by encoding as a finite state network all of the legal combinations of morphemes (morphotactics) and then subsequent rules can be applied as a finite state transducer. How this works is that a finite state transducer is created for each word and then the transducers are compiled together

to form one. Ideally, this should contain all of the morphemes, derivation, inflection, and compounding. (Finite State: a tutorial).

Finite State Machines have been used to analyze many of the European languages, including English, Spanish and German. In *Finite-State Parsing of German, Erhard Hinrichs* (2005) discusses why finite state morphology is interesting to implement for German, which he attributes to German's treatment of complex prenominal modifiers and also how finite state systems handle German clause structures. He touches on some of the different finite state software and approaches which can be used.

4.2 Xerox Finite State Software

At the time of the publishing of Hinrich's article about Finite State Morphology for German, no one had developed a finite state system using the finite state methodology Xerox software developed by Beeseley and Karttunen (2003). The Xerox Finite State Software has been used by the Xerox Research Centre Europe (XRCE) and the Palo Alto Research Center (PARC) to develop a variety of linguistic tools such as tokenizers, morphological analyzers and generators, POS taggers, and chunkers (Beeseley, 2003). An Online Spanish Analyzer/Generator was developed by Tinsley (Graham, 2007) in the XFST programming language of the Xerox Finite State Software. Actually, the complete program developed by Tinsley takes entire sentences and tokenizes them. Then each word is individually analyzed further. The surface form, or input, generates an output that is the lexical form and includes a series of tags that explain the surface form.

Surface Form: hablo Lexical Form: hablar+Verb+PresInd+1P+Sg
Figure 4.2 Surface form and Lexical form for two-level finite state morphology with Spanish example (Beesely, 2003).

Tinsley's product demonstrates the very practical purpose of morphological analyzers and how a finite state system is a useful approach.

The XEROX software is innovative in that it allows for finite state morphology to be applied to languages that have reduplication and templatic morphology that typically, under the simplest definition of a finite state system could not be incorporated in to the computational grammar. This has allowed for the inclusion of languages like Arabic, which are different from most of the Western European languages, in that has reduplication and templatic morphology elements inherent in their structure. Prior to the development of this software, it was not possible to describe such a language using finite state methods. For the flexibility and advanced capabilities of the Xerox finite state morphology software I will be creating a morphological analyzer of Mohawk with this program and using this method.

4.3 Corpus Based- Machine Learning/Statistical

However, the current trend that is at the forefront of NLP tools is the use of corpus based-machine learning approaches. The corpus based approach requires a large amount of written documentation in a given language that is used as a training corpus implementing a learning algorithm. The largest language corpuses have millions of words. Since the algorithm learns and uses statistics, therefore the greater the size of training material, the better accuracy the program will achieve. This is at the forefront of computational linguistics, however, for many of the world's languages, and for polysynthetic languages there is often not sufficient written documentation of a language for a training corpus to be created. In my attempts to create a morphological analyzer for Mohawk, I was able to gather close to 20,000 words, which might have been sufficient enough to attempt Machine Learning techniques. However, while 13,000 words were sentence aligned between English and Mohawk, the material was highly specific, taken from a Mohawk Language Standardization Project (Kanienkehaka.com, 2013). In comparison with the other Iroquoian Languages and with many other polysynthetic languages, Mohawk has a good size quantity of material, the Bible being the first translation in the 17th century. Yet, the spelling was not standardized and many resources in print utilized a computer to scan the original documents that changed letters and created separations within words. Thus for this project, a machine learning algorithm and technique was not attempted.

4.4 Morphological Analysis of North American Languages

Not only are the European languages or languages with large amounts of speakers becoming a focus of NLP applications development, but in the last 10 years a trend has been growing to utilize technology to help describe and develop support tools for the world's endangered languages. Every Native North American language spoken today across Canada and the United States is under threat, even Navajo, which has the largest population of 150,000 speakers (Ethnologue, 2013). This trend is coming from communities themselves who may have just one member who is a computer programmer and interested as a hobby to pitch in effort to save his or her community's language. Anthropological linguists are recognizing the value of these languages for the community, computational linguists are intrigued by the challenges they pose to NLP development, and government officials are starting to recognize and acknowledge the role they played in the accelerated extermination of the hundreds of languages once spoken on this continent.

In Canada, the government helped fund the development of a morphological analyzer for Inuktitut that as of April 2013 is now an official language in the Province of Nunavut, on equal footing with both English and French. This has been a major victory for the rights of the native North American languages. Researchers at the Institute of Information Technology of the National Research Council of Canada (NRC) developed a morphological parser with the mission to, "facilitate the use of Inuktitut in its written form on computers and the web by providing useful tools and links to important resources" (Inuktitutcomputing.ca, 2000). The creation of a morphological analyzer was key to fulfilling this mission because as a polysynthetic language Inuktitut can have words that, "grow to gigantic proportions, often counting over five or six infixes, and need to be broken down into units of meaning in order to be understood" (Inuktitutcomputing.ca, 2000). While not all Native North American languages are polysynthetic, there is great linguistic diversity and the language typology differs dramatically from the Indo-European languages.

A project to create a morphological parser for Lushootseed using finite state morphology using the PC-Kimmo software was one of the first finite state systems created for a polysynthetic

language (Lonsdale and Matsushita, n.d.). They were also able to use a Lushootseed dictionary that was already marked up in XML, which significantly eased their task. They knew that with the rich morphological inflection of Lushootseed that they would not be able to apply the Porter stemming algorithm, which is what is used for English search engines. This is why they utilized the two-level finite state morphology, thereby needing a lexicon, rule files, and grammar files. They created a separate lexicon for each of the possible positions where inflection and derivation could occur for a Lushootseed word. The word is processed letter by letter, and depending on which sub-lexicon the system goes through, it helps direct what sub-lexicons may follow later in the system. By the end of the project they ended up with more than 10 lexicons. It is important to add that within the PC-KIMMO program, only ASCII transcription can be used because UTF-8 was not handled by the program.

However, two of the most important works pertaining to this project about the Mohawk language come from Memorial University under the direction of Professor Carrie Dyck, who has worked with the Cayuga language as a linguist and with the community for supporting language learning and revitalization of Cayuga (Dyck and Kumar, 2012). As a fellow member of the Northern Iroquoian language family, Cayuga and Mohawk share many morphological similarities and characteristics. As a linguist Dr. Carrie Dyck has been publishing articles about the structure of Cayuga for years, and it was with a Master's student's thesis that they developed one of the first NLP tools for Cayuga (Graham, 2007). In 2007 Dougal Graham attempted to create a Finite-State Parsing system of Cayuga. He was the first person to attempt using finite state morphology for the Iroquoian languages, which are known to be some of the most highly inflected languages in the world. He only attempts to create a finite state machine that analyzes nouns, because the Cayuga verb structure, like Mohawk, is much more complicated and would be too great a scope for this thesis. He writes his program in XFST and LEXC using the Xerox Finite State software of Beesely and Kartunnen (2003). His findings are important for future development of NLP tools for Iroquoian languages because Graham's first attempt at creating a morphological parser for Cayuga was successful and the first major attempt to develop NLP tools for languages in this language family. He actually creates two

systems, one that is an abstract system, and one that is concrete in order to observe which approach functions the best. He also did this to see if it is possible to create rules for Cayuga that would ultimately make the program lighter and possibly run faster. The abstract version therefore requires more rules to be deduced and applied to the lexicon. It allows for fewer lexicons, whereas the concrete system does not rely on any rules to help with the parser except for simple spelling corrections, such as deleting a double vowel when it is created during the merging of two morphemes. Graham successfully builds both programs, and predicts that it would be possible to utilize finite state morphology to describe Cayuga nouns.

The next work completed was with Ranjeet Kumar, another Master's student, along with Dr. Carrie Dyck in *A grammar-driven bilingual digital dictionary for Cayuga (Iroquoian)* (Dyck and Kumar, 2012). This paper unfortunately did not explain which approach they used in the development of this program. However, they explain that it is an interdisciplinary approach using linguistic and grammatical resources along with computational data mining techniques. They address the inherent challenges in developing a bilingual dictionary that are also the same elements that make morphological analysis both necessary and difficult to perform, such as the word-internal grammar of Cayuga and all of the Iroquoian languages. It also includes handling the obligatory prefixes and the large variation in prefix structure. In the case of Cayuga they explain that there are over 50 prefixes to describe person and number, and to complicate a situation many prefixes have multiple variants (Dyck and Kumar, 2013). They also had to overcome the issue of orthography in Cayuga because Cayuga was not written down until the mid-20th century and the writing system is still under development and not standardized. They also had to decide whether to follow the phonemic or phonetic spelling of words, which can sometimes be quite different upon first glance. The audience of this product is also important to keep under consideration since it is a dictionary that is meant to be used by students of Cayuga. Most students prefer the phonetic spelling, however the language is written in a phonemic spelling. This phonemic spelling is the Henry Orthography. One solution was to include both the phonetic and phonemic lemmas in the output for learners to see not only how it is pronounced but how it would look too. The dictionary also included a feature that helped suggest the

word that the user was trying to input. Additionally, a really helpful feature for second language learners is that a student can find words that contain a partial match to their input.

This has been a successful product that is used by the Cayuga Nation and Cayuga language learners. As such a tool, this dictionary is on the Cayuga Our Oral Legacy website and is free for anyone to download (Hasler, 2013). However, when I input some known Cayuga words, no result came back, which could be due to differences in the orthography used for the dictionary and may also be because one needs to be given some instructions how to input more complicated words that do not consist of only the standard ASCII letters.

Having read about the projects for the Cayuga language I wanted to attempt to work toward developing a similar tool for the Mohawk Language, which has many more speakers and has been gaining momentum in its language revitalization efforts. Using finite state morphology I attempted to learn how to apply finite state morphology and the Xerox Finite State Morphology software using the xfst and lexc languages to create a simple program that can describe Mohawk Nouns. Following Beeseley and Karttunen's book (2003) I taught myself the xfst and lexc languages, which are the languages the finite state software, and wrote a program to handle Mohawk formal nouns based on a classification system that I developed from comparing three major language resources of Mohawk.

5. An Experiment in the Analysis of Mohawk Formal Nouns

5.1 Objectives

The objective of this experiment is to attempt to create a finite state morphological parser using the Xerox State Morphology software program that is bidirectional for Mohawk formal nouns, which is capable of breaking down these Mohawk nouns into their morphemes to identify the root of the noun and the role of the affixes attached to it. Additionally, though not a part of the primary goal is that the system could generate a legal Mohawk word as an output from an English input. Only Mohawk formal nouns were chosen for this system due to the relative uniformity of the structure and affixes along with a few exceptional nouns that have an unusually simple structure which does not include many affixes. Mohawk functional nouns were not included in this system because many of these nouns follow a structure that is far more complex and more closely follows the structure of Mohawk verbs because indeed many of the nouns in this category can, depending on the context, function either as a noun or a verb. Due to the highly complex nature of Mohawk verbs and time constraints for this project attempting to create a finite system to include them was not attempted.

This design and attempt to create this finite state system is not intended to be exhaustive of all Mohawk formal nouns in existence, but to be a system whose structure should be able to accommodate the majority of Mohawk formal nouns. For this reason only thirty English words, which results in thirty noun roots were chosen. Some of these roots will take the different nominal prefixes, suffixes, locative and attributive suffixes, along with possessive prefixes. The goal is to create a finite state system structure to which more noun roots and affixes can be inserted in to the program without having to make large programming adjustments. This system will not attempt to include the few affixes that require long distance dependencies to be incorporated into the finite state system requiring use of flag diacritics within the Xerox system.

5.2 The Lead Up to Mohawk and Finite State Morphology

Before deciding to work with the Mohawk language I researched what projects had been done and were currently underway with Native North American Languages and involving NLP. I looked in to working with Inuktitut, Michif, Seneca, Cherokee, Cree, Ojibwe, Mohawk, and Navajo. I discovered that the Canadian government had already developed a morphological analyzer and was developing tools needed to launch a search engine in Inuktitut. This was a great inspiration for my hopes to begin working toward the development of a Mohawk morphological analyzer. I reached out to the coordinator of the Michif Dictionary Project hoping to become involved in some way with this language that is truly a unique linguistic development, being a combination of Cree, Ojibwe, and French. It is not quite a Creole, but a true mixture of these languages that was spoken by the Metis People in Canada, which used to be a large and thriving culture. The Metis were people that were mixed of European, mostly French, and indigenous peoples, most commonly Ojibwe and Cree. However, this language was not polysynthetic, and there was currently no need to me to become involved except for data entry of words.

After looking into Michif, I then considered Cherokee as it is one of the more famous North American Tribes, for the Trail of Tears in the 1800s, and Cherokee has the most speakers of any Iroquoian language. Work with this language is more developed and 2013 saw the release of a Cherokee video game by Thornton Media, Inc. However there were many projects currently underway for Cherokee, including versions of digital dictionaries. This led me to The Cree language, which I had been familiar with from experiences dogsledding in northern New England and Quebec, Canada. Since Cree is one of only three First Nation's languages that Statistics Canada predict will survive to the next century it became not the best option for research since there already exists a team of people dedicated to developing a morphological parser and other NLP tools. Then there was a quick consideration of Ojibwe, one of the other three predicted to survive, which has NLP tools and iPhone and android applications already developed and available on the market. The Navajo language has the most speakers of any Native North American language and has been the subject of extensive linguistic research. Moreover, NLP tools, including a language pack for Microsoft Windows have

been developed for it. Additionally, though not related to NLP, Star Wars in 2013 was dubbed in to Navajo (Jones, 2013).

Many of the individuals or organizations I reached out to either did not respond, or ultimately were unsure how I could be a part of their already existing project. Through an article I found from the Rochester Institute of Technology on their university website about collaboration through a grant between the university and the Seneca Nation in western New York State, I got in touch with a Professor that was coordinating the development of a Seneca online language dictionary. I was able to speak with the primary developer, Robbie Jimerson, who was an MSc graduate in computer science from RIT and also a member of the Seneca Nation. He is a recent graduate of the program and had become interested in working to preserve and revitalize the Seneca language out of personal interest. His grandfather spoke Seneca fluently, but between his grandfather's generation and his generation there has been almost no language transmission (Jimerson, 2013). Jimerson, not fluent in Seneca, then teamed up with Dr. Wallace Chafe, who is the leading expert on the Seneca language and was one of the first linguists to document and describe the Seneca language in the early 1960s.

Dr. Chafe published in a series of eight articles description of Seneca Morphology in the International Journal of American Linguistics in 1960-1961. He has also worked with and performed historical linguistic research concerning the Iroquoian language family. Much of his historical research included Mohawk since the Mohawk language is one of the Native North American languages with the earliest documentation. From his research he deduced that contemporary Mohawk is most directly related to Iroquoian, a language that at one time was the one language all the tribes spoke, and that as one travels further west across New York State that the languages experience more innovation and change in syntax and phonetics. The Seneca language, being the language of the tribe known as the western gate-keepers was the farthest geographically from the territories of the Mohawk and thus he predicts this is why Seneca experienced the most radical phonological changes. This is important to mention because Robbie Jimerson during a conversation about the Seneca Language Dictionary Project explained that in order to attempt a digital bilingual Seneca-English Dictionary that Dr. Chafe had had to develop a method that included passing the word through 17th century Mohawk,

due to the phonetic degradation that has occurred in the Seneca Language. Often the root of the word is no longer pronounced in Seneca, even though speakers understand the meaning. Thereby, in Dr. Chafe's system a Seneca word as input is then converted to 17th century Mohawk, which allows for a root to be inserted into the Seneca word. This then makes a translation possible into English. When going from English to Seneca the English word is converted in to 17th century Mohawk, and then in many occurrences the root must be deleted when making the final conversion in to Seneca. Robbie was unsure of what methodology or grammar formalism was being utilized but he did not believe it to be finite state morphology. However, despite attempts to contact Dr. Chafe, no response was ever received. The site for their project can be visited at www.senecadictionary.com. Robbie was enthusiastic to have some help with the project, but only as data entry.

After having a positive experience with Robbie about Seneca, it seemed interesting that people were dedicating so much effort to languages with under 100 speakers, such as Cayuga with approximately 60 and Seneca with fewer. However, it seemed that no one was working toward a product for Mohawk. In 2012 Monica Peters, an iPhone application developer and a member of the Mohawk Nation did develop an iPhone application released to help people learn the Mohawk language, but it only provided some basic word translations (Talkmohawk.com, 2013). Considering the well documented difficulty of print bilingual dictionaries for English-Mohawk due to the morphological inflection along with the tremendous efforts and successes of the Mohawk communities to increase language learning, it seemed that this would be a well-timed project.

5.3 Resources, Tools, and Materials

After reaching out to those working on NLP tools for various Native North American languages and performing research on available resources for these languages, because I was not going to have either the time or the access to speakers or experts of any of these languages at my disposal in Ireland, I decided to move forward with Mohawk. It had a need for work to begin for a dictionary project and also had written documentation in far greater quantity than many other Native American languages. Once this decision was made it was imperative to collect every resource possible

in the Mohawk language. With the help of the internet, Amazon's website and access to other online materials I was able to get print copies of most materials in print in Mohawk for learning Mohawk. One of the first resources I gathered was *Radical words of the Mohawk language, with their derivatives,* which was a reprint of the 16th Annual Report by the State of New York, in which included even older works that described some of the earliest grammatical and lexicographical works ever created of the Mohawk language, which was gathered by Father Jacques Bruyas in the late 1600s. Unfortunately his dictionary is French-Mohawk bilingual and therefore proved of little use to me, French not being one of the languages I can speak or understand very well.

Another source of Mohawk documents came from a company BiblioLife. I was able to order the *Acts of the Apostles, in the Mohawk Language,* which is a part of their reprint series and is a scan and reproduction from 1835 translated by H.A. Hill and printed by the Young Men's Bible Society of New York. They also did have *the Book of Ruth* and other chapters of the Bible available in Mohawk. However, while it was good content and a high word count, in order to create the reprints the books had been scanned and during this process Mohawk words were broken up unnaturally, and the scanner did not have optimal accuracy in reading all of the original document's letters which has led to changes that make the content inaccurate. This then became another resource that I had been hoping might be useful to attempt to apply a machine learning algorithm, but was ultimately not able to use. Another resource was from the American Language Reprints that was that *Early Vocabularies of Mohawk.* This resource was an English-Mohawk and Mohawk-English dictionary along with sections for Mohawk numerals and so on. This reprint included works by the Dutch settlers in the first half of the 17th century who would have been the first European community to meet the Iroquois Nation and the Mohawk. While this resource was potentially useful, it included no explanation or classification of Mohawk nouns, and due to the highly contextual nature of Mohawk that affects the surface form of a noun it could not be used with certainty because I do not have an in-depth knowledge of Mohawk myself as a speaker to understand the differences.

The most useful language resources were the PhD thesis of Mohawk grammar, and then teaching tools that have been developed and improved since the 1970s (Bonvillain, 1973). Dr. Nancy

Bonvillain in her 1973 doctoral dissertation spent a year creating *A Grammar of Akwesasne Mohawk*, which was the first description of contemporary Mohawk. I was planning to model my finite state system only based on her classifications of Mohawk nouns and verbs; however, while it was an excellent description of Mohawk, it was simplistic and only described the language spoken in the Akwesasne community. Additionally in the early 2000s the Mohawk community set out to standardize the orthography of the Mohawk language and the differences in variation between the reservations, which couldn't have been addressed by Bonvillain at the time she was performing her research. This being said, her dissertation was the first resource that broke down and spelled out the different ways to classify nouns and verbs based on different affixes, dealt with some of the semantic classifications, and provided spelling rules for combining different morphemes. She explained that there were two types of nouns, formal and functional. In the formal noun category she claimed that there are only two nominal prefixes, and included many words under a classification that did not follow the formal or functional noun scheme. Had I followed only her dissertation or had only her resource available then my project would have been a simpler affair. It is important to add that for the purposes of her dissertation she was not performing a morphosyntactical description of Mohawk but a phonetic study, which is perhaps why some of her morphotactic descriptions have been revealed to have shortcomings and not be complete.

However, more resources have since been developed for Mohawk that provide a more comprehensive description and breakdown of the classifications, even though they are not intended for linguists but for language learners. Upon scanning these resources it was very quickly confirmed that due to the high complexity and variation in the composition of Mohawk verbs that creating a finite machine to describe Mohawk verbs would be too great an undertaking considering the scope and timeframe available and my lack of background in formal linguistics, computer programming, and the Xerox Finite State Morphology software. Since Mohawk functional nouns are really structurally similar to verbs, I also decided that it would not be possible to undertake a project that would both describe formal and functional nouns. This left me with the challenge to create a system that would describe the vast majority of Mohawk structural or formal nouns.

With this decision in mind, I then began investigating what resources existed for studying Mohawk that would provide any explanation about Mohawk nouns and how to form them. One of the first resources I found was online in a PDF form, from the province of Ontario, Canada they have a curriculum resource for grades 1-12, *Native Languages A Support Document for the teaching of Language Patterns* that explains the grammar of Cayuga, Oneida, and Mohawk extensively (Edu.gov.on.ca, 1999). It walks readers through the different parts of speech, Particles, Nouns, and Verbs, explaining the different classifications and giving examples of the different affixes in each language. It is a large resource at 144 pages. They present four types of nouns: structural, nouns with animate prefixes, unanalyzable nouns, and verbal nouns. The verbal nouns in this guide coincide with the functional nouns category in Bonvillain's dissertation. Additionally, both really do have an unanalyzable nouns category. However, this resource acknowledges the presence of nouns with animate prefixes. This distinction is crucial because this is a complex category that combines nouns that follow a structure similar to structural nouns and also nouns that are structurally more like verbs but semantically are terms for people and kinship. Bonvillain actually included kinship and people nouns in her verb section, even though they do not follow strictly the verb structure either. The Native Languages resource guide not only includes a more comprehensive description of Mohawk, Oneida, and Cayuga, but also explains that in many cases, the different prefixes for structural nouns can be based on the noun's attribute, whether it describes an object in nature or is man-made. This type of classification would help support the use of a concrete finite state machine because nouns would need to be added to their semantic classification. A possible situation would have been if the pronominal prefixes were determined more on the first letter or phoneme in the noun root, but this is not reliably the case.

Another resource also discusses this distinction, *Kanyen'keha Tewatati (Let's Speak Mohawk),* by David Maracle (1993) where he makes reference to nouns that describe objects that are man-made or that have been influenced by humans to have the prefixes ka, a, and rarely e or en, whereas nouns that describe objects found in nature that are not influenced or changed by humans take a different prefixes of o- or on- (1993). He chooses to break down the nouns into two categories

of formal and functional. However, within functional nouns, that have a structure similar to verbs, he includes two further categories, and in the second category he places nouns that describe people and kinship terms due to their mixed structure of formal and functional. In his other book *One Thousand Useful Mohawk Words* I was hoping to essentially find a long list of words that would include formal nouns that I could insert into my finite state system, but most of the words were phrases, or functional nouns, which supports comments made by many linguists of the Iroquoian languages that they are very verb heavy (1992). Thus it was from his *Let's Speak Mohawk* that unexpectedly I found the most succinct and laid out explanation for formal nouns and their affixes.

The last resource that I utilized was *Mohawk A Teaching Grammar* by Nora Deering and Helga Harries-Delisle (2007), which comes from the Kahnawake Reservation outside of Montreal. This is an incredible resource with by far the most amount of words listed and providing good detail about how to construct words in Mohawk. However, lists of words were in part broken down by prefixes and classifications similar to David Maracle's, but the material was also presented thematically, which would sometimes include multiple categories of nouns that would include the varying formal and functional nouns. Since I am not a speaker of Mohawk this made it more difficult, but not impossible to sort out the words that I was trying to include in my finite state system. Additionally, whereas Nora Deering's book was created and printed in Montreal by the Kahnawake Community, David Maracle is from the Six Nations territory in eastern Ontario. This in consequence included some dialectal variances. What I was ultimately hoping to do was to utilize elements of Bonvillain's dissertation, Maracle's, and Deering's books in order to cross reference each other and to incorporate the different dialects and spelling considerations into my system. However, upon making this attempt it became clear that I did not have the tools, nor the background to truly develop a full linguistic plan that would incorporate all three of these with any ease. The product of my work is therefore an attempt to create as complete a linguistic plan intelligently and correctly incorporating the three sources and their schematics for Mohawk formal nouns in to one break down in my finite state system.

5.4 Methods

5.4.1 Learning Finite State Morphology: Xerox Software and lexc

After gathering all of the practical linguistic resources possible of Mohawk I then turned to Kenneth Beesley and Lauri Karttunen's book *Finite State Morphology* (2003), which also included a CD to install the Xerox software on my computer, because I had never worked or studied finite state systems before and the Xerox software has been widespread in its use by linguists developing finite state machines for languages. Before any work could begin working toward a system for Mohawk I had to work through the book and teach myself the basics not only of finite state morphology, the concept of which I was familiar with and that at least resembled some similar philosophical design elements to my Localisation Process Automation module. At the start of this book I had no prior computer programming experience, but was not deterred by the warning in the introduction that realistically one really ought to have C++ or some previous experience with a computer programming language in order to understand the design of how both xfst and lexc languages work. The following paragraphs describe my process of learning the Xerox software and the xfst and lexc languages by working through the Beesely and Karttunen book chapter by chapter and how this process aided me in the development of my own lexicon in lexc for Mohawk formal nouns.

In Chapter 1 I worked through the beginning and basics of finite state morphology including terminology and how ultimately a lexicon and a set of rules are created for a finite state machine that then in conjunction work together for the final product, one of which can be a morphological analyzer and generator. For the first time I was exposed to simple regular expressions in Chapter 2, and learned about the different functions I might be incorporating in to my system such as union, concatenation, and composition. It was in this chapter that I began working with the xfst language and program and working through the basic exercises. In order to follow along I used Notepad to write up any code I needed to write to complete the exercises and then learned how to insert the code into the Xerox xfst command window. Chapter 3 is dedicated to the xfst interface and I spent a lot of time in this chapter learning my way around the program and doing the exercises for the languages they provided.

However, this chapter was overwhelming and I was unable to resolve some of the exercises nor perform the advanced xfst commands.

I did not let that stop me from continuing on the Chapter 4 because this is where they present the lexc language, which is the language in which I have written my program for Mohawk entirely. The lexc language is much more straight forward, and visually easier to follow because it consists of setting up categories or lexicons and sub-lexicons. Immediately below the title of the Lexicon or sub lexicon is a list of the morphemes that pertain to it, and then on the right side is a marker that indicated what is the next sub lexicon may follow each specific morpheme in that sub lexicon. It was while in Chapter 4 that I realized that if planning to develop a concrete finite state system for Mohawk that I would mostly be using lexc, and that also considering the simplicity in its design in both the organizational layout and to allow for someone like myself with no programming experience to understand it, that it was where I needed to place my focus.

Chapter 5 was about planning and taking linguistic strategies for planning finite state projects. This chapter I was hoping would help me understand how to integrate an xfst rule program with a lexc program that included the bulk of the lexicon by working through the exercises provided. My theory was that it would help me gain knowledge about how they resolved the linguistic complexities of languages using lexc and xfst, but I found this not to be where the solution was held. However, it did walk me through flag diacritics and how to handle long distance dependencies in a finite state system, which is a function that allows finite state systems to handle this element that normally cannot be handled. After Chapter 5 I did not continue with further chapters, except to read up more on flag diacritics in Chapter 7. Upon reading this chapter and attempting to understand how to practically apply it I decided that I would not be including the few cases of nouns that would utilize this function. I did make an attempt, but the few nouns that include this are certain plurals, and certain number nouns. To work through Beeseley and Karttunen's book took an entire month of full-time focus and study because I was starting with no previous background in either linguistic planning, computer programming, nor in finite state morphology.

5.4.2 Linguistic planning and execution

Having finished working through Beeseley and Karttunen's book, I then decided to begin working on the lexc document that would become my main product keeping in mind the three Mohawk language resources and how they organised the prefixes, noun roots, and suffixes. The letters used in the Mohawk alphabet were all used except for the glottal stop, and diacritic marks along with the vowel length marker were not included in the system. I also developed a multi-character set of symbols, following closely the examples from Beesely and Karttunen and Graham's thesis with Cayuga that included 16 multi-character symbols that would ultimately be an output for the system. Additionally I created names for the varying sub lexicons and divisions that I created in my program. At the very top I listed the multi-character symbols. *See table 5.1*.

+nomPF	Nominal prefix
+nomSF	Nominal suffix
+possPF1p	Possessive prefix first person
+possPF1p2P	Possessive prefix first person, two people
+possPF1pP	Possessive prefix first person, we
+possPF2pS	Possessive prefix second person singular
+possPF2p2P	Possessive prefix second person, two people
+possPF2pP	Possessive prefix second person, more than two
+possPF3pmS	Possessive prefix third person male singular
+possPF3pfS	Possessive prefix third person female singular
+possPF3pnS	Possessive prefix third person neuter singular
+possPF3pmP	Possessive prefix third person plural male
+possPF3pfP	Possessive prefix third person female
+locSF1	Locative suffix 1
+PLAn	Plural animate suffix
+PLInam	Plural inanimate suffix

Table 5.1 Multi-character symbols developed for the Xerox finite state software in lexc, for use with Mohawk.

This list included the basic +nomPf (basic nominal prefix) and +nomSf (basic nominal suffix) and then included all of the different possessive prefixes from singular to a symbol that represents plural with two people.

5.4.3 Defective Nouns and Pronominal Prefixes

The top lexicon I entitled 'ROOT' and then from there divided the ROOT lexicon into basicNouns and defectiveNouns. The sub-lexicon defectiveNouns would be simple nouns that did not

follow the structure of formal nouns, and that often did not take the same or many affixes. This was the first sub lexicon I developed in order to verify that indeed I was able to design a lexc program that would create a useful output. As I was writing the program, after each word I input into the defectiveNoun sub-lexicon I did not initially include a continuation that lead to the NSF sub-lexicon, but rather initially just put '#' , which is the symbol in lexc that tells the program during run-time to stop at that point. I did this so I could immediately see results when running my code to test that indeed I was creating a viable product. Organizationally, I created sub-lexicons for all of the prefixes, from there the system led to the noun roots, and the program ended with all of the suffix possibilities. For this reason I initially closed the words in the defectiveNouns category with the end symbol rather than leading to the nominal suffix sub-lexicon (NSFs) until I had reached the stage in my program's creation where I was inputting suffixes. Within the defectiveNouns sub-lexicon there was a list of the words in Mohawk, at first without the English translation. The next sub-lexicon called in this string was NSFs, which stood for nominal suffixes, because while these nouns did not take the usual prefixes, they could all take locative and attributive suffixes.

5.4.4 Basic Nouns and Noun Roots

In the basicNouns sub lexicon I included a fork to include another break down that would lead to either unpossessive or possessive prenominal prefixes. In the sub lexicon BasicUnPossPF I created three unpossessive prenominal prefixes: ka-, o-, aw-. The BasicPossPf sub-lexicon included the fifteen prefixes used in Mohawk to denote possession based on the number and gender of who possesses the object. These possessive pronouns describe and belong to alienable objects rather than inalienable objects, which would be body parts, for example. Inalienable objects take a different set of possessive pronouns. All of the prefixes whether possessive or unpossessive then lead in to either the BasicAllRoots, or directly to one of the specific noun stem sub-lexicons that was devised based on the first letter or two of the noun stem. The options were basicKaRoots, basicORoots, basicARoots, basicEroots.

Initially I had each prefix leading directly to its specific sub-lexicon of a basic noun root because the unpossessive prefixes only take one specific noun root except for the null pronominal prefix that can take stems in the basicARoots or basicERoots sub-lexicon. However, when I began including the possessive prefixes, some of the prefixes do not vary depending on the first letter of the noun stem and are applied generally to all of the noun stems. Rather than listing each sub-lexicon individually, I decided to include another layer and sub-lexicon of basicAllRoots that would include and make possible all of the noun stems in my system in all of those separate sub-lexicons. The noun root options are basicKaRoots, basicARoots, basicERoots, basicORoots. In each of these sub-lexicons I include a list of the Mohawk words, without their translation included initially. The path through this system then leads to a sub-lexicon that includes all of the suffixes including a null suffix.

5.4.5 Suffixes

The nominal suffixes lexicon was denoted 'NSFs' and leads to further sub lexicons to include attributive suffixes (attSF), locative suffixes (locSF), and the basic suffixes (basicSF). The attributive suffix sub lexicon includes the suffixes that would in English be an adjective. This includes suffixes that represent the English adjective equivalent of large, good, bad, old, new, and also includes plural animate and plural inanimate. The locative suffixes at this level lead to a possible morpheme that –atsher- depending on the noun root can be inserted between the actual locative suffix, sometimes in order to emphasize the particular suffix, and sometimes because the noun generally includes this extra morpheme between the noun root and locative suffix. Therefore in the sub lexicon of locSF it includes the two possible morphemes that might be inserted, and also a null morpheme, all three of which lead to the locfinalSF sub-lexicon which represents what in English would be expressed through the prepositions on, in, under, near. From these options I then use the # symbol to close the path.

5.4.6 Bilingual Mohawk to English and Testing

Only once all the sub lexicons had been created did I then go back and input the English translation of the prefixes, noun stems, and suffixes. However, the unpossessive basic prefixes do not have an English translation, which is why I used the multi-character symbols that I had created at the

beginning of the process. I knew I would be needing them and had included those symbols above the start of the lexicon ROOT to be possible outputs of the system in addition to the Mohawk and English words. Additionally, I used the multi-character symbols for the possessive prefixes because in English we simply do not have a one word possessive pronoun beyond *my, your, his, her, our,* and *their*. It is not easy for us to express *our (two of us)* in English and for this reason, rather than alternate between the English translation for the pronouns we do have and multi-character symbols I decided to include all multi-character symbols.

To test my program I input various combinations of the Mohawk morphemes under the 'apply down' command. I chose combinations of prefixes roots and suffixes that would make sure all of the sub-lexicons would be called on. If the combination was a valid option in my finite system then the output would be the values of translation of those morphemes in English. Figure 5.1 yet to be numbered, has a list of morphemes, including combinations that were not legal in the system and do not have any output.

Figure 5.1 screen shot of Xerox software running the lexc program for Mohawk formal nouns. Examples show the input and output breakdown of the morphemes.

I also tested the system in the reverse direction since a finite system should work bi-directionally. This involved inputting only the English word, which would be the equivalent of the noun root and also inserting the nominal prefix symbols and the nominal suffix symbols around the English word to see what the system would output. Figure 5.2 shows that using 'apply down' that the

system can return the only the noun root when the input is the basic English word, and the system can also output the complete Mohawk word if you call for the prefix and suffix included.

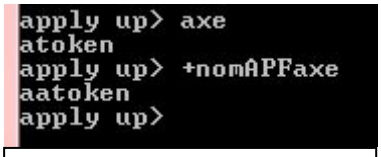

Figure 5.2 demonstrating the possible inputs and output capability of the program.

5.5 Results

5.5.1 The system and organizational structure

The final product of the finite state system is written in lexc and run using the Xerox Finite State Morphology software from Beeseley and Karttunen (2003). The finite state system is a morphological analyzer that can handle formal Mohawk nouns of a limited scope and structure. The analyzer/generator only includes a lexicon code and does not include a rules component that would help with rewrite and replace rules. To see the full code of the program please see appendix. The Xerox Finite State Morphology software successfully builds the Mohawk lexicon and after minimization creates a system that is 9.7Kb, with 237 states, 358 arcs, and 8,215 possible paths. To have 8,215 possible paths, or outputs it took only 30 noun roots, 19 prefixes, and 20 suffixes. There is both a null prefix and null suffix as valid inputs, including a null morpheme suffix that may or may not precede the locative suffix ending. Aside from the possible double suffixes for locative affixes, the structure of the program only allows for one prefix, one noun root, and one suffix. The analyzer goes from English to Mohawk and Mohawk to English. Since there is currently no rule component in addition to the lexc lexicon code, there are some possible paths are not legal outcomes. This is because the system was designed to be more inclusive rather than exclusive.

The defectiveNoun sub-lexicon are nouns that do not take any of the prefixes I included in my system but can have suffixes attached to them. In the basicNouns, the noun roots that are alienable objects have the ability to be input as either possessed or unpossessed. Almost all of the noun roots in

the system are alienable except for ‘face’ and ‘leg’. More inalienable nouns were not included because they take a different set of possessive prefixes, and follow a completely different structure of morphemes. I chose to include these two inalienable noun roots because they follow the formal noun structure all except for the possessive prefixes, and are a good example to demonstrate the complexity of Mohawk and some of the hurdles facing those who are working to develop NLP tools for this language. Another factor that relates to the possessive prefixes and when they can be applied is that they cannot be applied to animate objects either. Most of the animate objects included in the program are actually categorized in the defectiveNoun sub-lexicon, but some animate nouns such as *dog*, for instance, are listed under the ERoot sub lexicon. This noun root and others like it cannot be made plural simply by adding either an alienable or inalienable possessive prefix to the noun root, since in Mohawk it is impossible for a human to possess another living creature. Thus when attempting to express theoretical possession of an animal there are multiple and varying ways using verbs and other constructions that Mohawk speakers must use. Additionally, when adding a pluralizing suffix, whether the noun is animate or inanimate affects which plural suffix it will take. The noun roots were categorized based on the prenominal prefix they take, which was more or less straight forward but not by the characteristic of being a living being or not. The two plural suffixes are included in this system. The plural suffix for inanimate nouns is –okon, and for animate nouns is –okonha. Even though there are two plural suffixes, the system marks one as inanimate or animate so the user can input in English dog+PLAn and the output would be *erharokon* (dogs). However, if a user put in *erharokonha*, the output would technically be valid and would return dog+PLInan. The idea is that the user would understand that it is not a valid output themselves because a dog as an animate object cannot be made plural using an inanimate plural suffix. Most of the words in the basicNoun sub-lexicon are inanimate but there are a few instances where the animate plural suffix would be used and for this reason the animate suffix was included.

Since some nouns that begin with ‘e’ and ‘a’ do not take a pronominal prefix and whereas some noun roots that begin with e and a do take aw- this is why the null prefix was included in the basicUnPossPF sub-lexicon. However, this limits the program, since some noun roots beginning with

'a' do also take the prefix a- and this addition results in an output with aa- as the leading letters of the word, but Mohawk speakers would simply write the word deleting one 'a' or writing the word as a'. In this regard the program is not completely representative in high accuracy when it includes the pronominal prefix a- and a noun roots that begin with 'a', which is another evaluation issue. In order to increase accuracy, a rewrite rule would have to have been created in xfst code within the rules program and since the spelling change of either aa- or a' does not follow any rule, it would have had to have been entered specifically for each noun root. Thus the morphemes can be added to one another but the program does not change or address any spelling changes that might need to be made. However, it appears there are few cases in Mohawk formal nouns where this occurs so it would not necessarily result in a significant increase in performance of the system. Additionally, many of the spelling changes cannot be easily be captured by rules and then applied uniformly; instead, each combination of prefix-noun root, noun root-suffix, suffix-suffix needs to be treated as a special case. Moreover, the way in which such a combination is treated can vary across dialectal regions of Mohawk.

For the suffixes, the locative suffixes were the most complicated to add because some nouns take an additional morpheme that goes in between the noun root and the locative suffix. It would also be difficult to write a rule for this action and when it occurs because Mohawk language teachers claim it to be arbitrary and each noun root that takes the intermediary morpheme must be learned individually. Additionally, some noun roots take this intermediary locative suffix to change and manipulate the meaning. Another hindrance to attempting to create a rule o in xfst to capture the cases when specifically this morpheme can be inserted is that it also varies per regional dialect, with communities of Akwesasne vs. St. Regis, Kahnawake, and Six Nations each utilising it for different words and with a different semantic meaning. A similar and yet different situation is the attributive suffix that in English would be represented as the adjective *big* does pose a problem because there are two suffixes that correspond to the English *big* and which noun root takes which suffix is completely arbitrary. *See figure 5.3.*

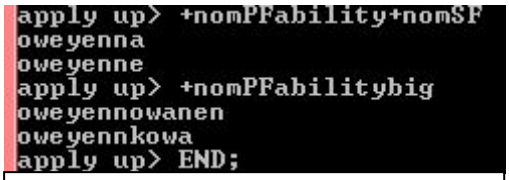

Figure 5.3 demonstrating how to use the Xerox program with my lexc code with an English input calling for both a prefix and suffix either nominal or attributive.

Additionally, much like the locative suffixes, some noun roots can take an attributive suffix that is usually a regional dialectal preference, and as such that particular choice of one suffix versus another may have certain attributes with it that subtly change the meaning. The suffix –owanen is the most commonly used suffix to express 'big', but –kowa can be used as well though typically in the Six Nations Reserve, it is a suffix used when expressing that the object is "big" or large for its kind, or for what it is (Maracle, 1993). Since these two can be used either interchangeably, or for slightly different intentions, both outputs are considered valid, but the user of this program will need to understand each suffix, or need to clarify with another source the difference in meaning for each.

5.5.2 English to Mohawk, and Mohawk to English

When going from Mohawk to English the output does not explain which portions of the word are the prefix, root, and suffix, and the unpossessive prefix output will for any of the prefixes ka, a, o, or aw not tell the user which prefix was used. It would seem then that a user of this program would need to know the prefixes for the words. Additionally, the basic suffix ending output will include whether it was /a/ or /e/. From this output even though the return for the noun root will be the English translation, the user can infer the noun root. None of the prefixes or suffixes include in the output what letters specifically belong to them. It was not included because they were already longer multi-character symbols and I did not want the return to be so long, which might possibly create more confusion and difficulty in locating the noun root. Additionally while when going to Mohawk from English this would be helpful for a user with little knowledge of the language, it would make it difficult for users to go from English to Mohawk because then in this case they would need to already know the prefix in order to call the correct word, thereby making the program less user friendly for

non-speakers or beginners. This being said, it is completely possible using only lexc, to include the 1-4 letters that corresponded with the Mohawk word in the return along with the English output.

An interesting result is that when going from English to Mohawk, for some nouns it is possible to not include the call for a nominal prefix nor a nominal suffix, and yet the system has a valid output. *See figure 5.2 on page 52.* In the Mohawk language, at first glance it would appear that neither of these outcomes is valid because in the first input the inclusion of a nominal prefix was not called for and the system gives an output, therefore verifying that it is legal in the system and then both would be illegal because the output does not call for a suffix ending. If we consider the situation with the prefix first, then recall that the code allows for both roots that may take the a- prefix and roots that may take the aw- prefix, that are considered ERoots, may also appear without a prefix in certain circumstances. To have avoided the *aatoken* output, removing the a- prefix would have solved the spelling error that was created by the output of +nomAPF with an A-root noun that begins with an a, but I then would have had to have included the a with the noun root always, which would have been an error in assignment because in some circumstances there is an /a/ in the root, but in many it is only the prefix, and in some there is a leading /a/ in the noun root and additionally an /a/ in the prefix, which results in a spelling change of either deleting the extra /a/, or adding an apostrophe after the one remaining /a/. Since I did not create an xfst rules program to accompany the lexicon in lexc, I chose to leave the outcome without the spelling modification. However, a user of the system if this program was functioning as a bilingual e-dictionary, would need to already understand that this was meant to be corrected and would need to have a Mohawk speaker available to help with outputs like this.

The current system is not very strict with following Mohawk grammar rules for constructing formal nouns, and thus there are some outputs that are not legal in the Mohawk language. However, with the hopes to be more inclusive than exclusive, the system will go from English to Mohawk and vice versa and has an interesting feature that allow users to input only the noun in English without calling for the prefixes or suffixes, which would give the user an output that is only the noun root. The system created can function as a rudimentary bilingual e-dictionary for Mohawk language learners

and also is a starting foundation to create a morphological parser, which could be used as a tool to develop more advanced linguistic tools for Mohawk.

6. Discussion

6.1 Analysis of the System

The program created for Mohawk formal nouns is in many ways very simplistic and incomplete, since it only includes 30 noun roots, however even with such a small number, and the limited type of nouns that this system can describe, with those roots and varying affixes one can appreciate the complexity and flexibility inherent in the Mohawk language word structure. Since the program does not include a rules component that would be applied in succession after the lexicon, there are certain spelling corrections that are currently not corrected, and without the rules component it allows for increased flexibility within the structure that is laid out in the lexicon, but is also limiting. For instance, if a user inputs just the noun root, then the translation of it without its required affixes is legal. This is potentially a good quality because it allows a user, if they are familiar with the program and how it works, to actually locate just the noun root, but yet it is also misleading because that output is not a stand-alone word in Mohawk and indeed must have certain affixes in order to be considered a legal word. In a rules code, written in xfst, the developer could write rules for the structure of the output that is must include a prefix, noun root, and suffix. This could allow users to input only the English word and the result would be a word in Mohawk that included the noun root along with its basic nominal prefix and suffix. This still could allow for outcomes that might not be legal, because both a null prefix and null suffix are defined in the lexicon to allow for the words that do not take either or both, but perhaps this could be adjusted for in the rules component as well. Designing this system was not incredibly difficult up to this point and with more linguistic understanding of Mohawk, and computer programming skills it seems likely that much more progress could be made toward the development of a morphological parser of Mohawk using finite state techniques.

In regards to the structure of Mohawk words, what has not been discussed is anything beyond the prefix-root-suffix structure, to include multiple suffixes. Indeed it is possible to include locative,

attributive, and plural suffixes all on the end of one noun root. Within the lexicon I believe it could have been organized in such a way to allow for multiple suffixes to be added to one another, but there was little information I found regarding how to organize and order these morphemes from the Mohawk language resources that were available. Certainly any allowance of this within the structure of the lexicon would be more inclusive than exclusive and would not be able to give a true representation of the order in which the morphemes need to be placed and any other structural restriction that could be necessary. In order to do this it would have essentially meant building in a loop into the system, and for the reason not being too inclusive and possible never-ending, I decided not to include that loop structure that would allow for multiple suffixes. I was also not how to organize it in a way to prevent multiples of the same suffix to be added. Even though this realistically would not be an input someone would attempt if they were using a Mohawk word from a real source, I knew it would make the program size much larger and did want that type of loop in the program. In order to make this a possibility I believe having an xfst rule program absolutely necessary and I first need to learn the exact hierarchy of the suffixes in the order they can be added. Plus, this new addition of multiple morphemes could also include more instances for spelling corrections and revisions after the combination of these morphemes. These additional spelling rules and the hierarchy of adding morphemes was not available from the resources I had available to me and moving forward would involve working with speakers of the language, and most likely teachers, who would be better equipped to explain the morphotactics of the Mohawk word.

6.2 Usability

The current product is usable and does create an output that is valid, but would require a user to understand many multi-character variables that are possible outputs for the different affixes. Additionally, one would have to already understand some of the basic structure of the Mohawk language if one was hoping to go from English to Mohawk. For instance if a user input ‘big house’ then the current system would produce no output because it cannot take the adjective in English as a separate word and appearing before the noun, which in Mohawk is a noun root, and convert its English counterpart to the Mohawk equivalent with not only a different order, but combining the

words to become one single word in Mohawk. In this case, the user must understand where that adjectives, preposition, or possessive pronoun in English, represented as separate words, fit in to the Mohawk word structure. For example a user would have to know to input 'housebig', which would then still give two outputs, whether the house is just a house that is big, or a house that is big compared to other houses. Another dilemma is whether or not to include in the +nomPF, the prefix itself, such as +nomKaPF, for example, that would infinitely help users going from Mohawk to English learn to understand what are the prefixes and further ease the task of identifying the letters that consist of the noun root. On the other hand, it makes it next to impossible for someone trying to go from English to Mohawk if they don't already know it because the program will only deliver an output if the input is correct and they call for the correct multi-character symbol prefix. Looking at the attributive suffixes, there are two suffixes for 'big' but there is no way in the system to express small or little, and this is because it is a construction that involves an additional prefix and a suffix. I was not able to work through how to incorporate a long distance dependency using the flag diacritic system of the Xerox Finite State Morphology. It is interesting to note, however that expressing large size is simply a suffix, but expressing a small size is more complicated in Mohawk, and much more complicated to resolve computationally.

Another aspect of this product is that currently it only runs through the Xerox Finite State Morphology Software in a command window, but the user interface is incredibly simplistic. It is also a program that as is, needs to be installed on one's computer in order to use it. Perhaps a possibility, much like the Cayuga digital dictionary would be where it is a free download from a website. However, the world of technology and language tools is moving away from this and toward internet based programs. Word Reference, and even the Inuktitut morphological analyzer is internet based. Then in this case, some institution or organization would need it to be loaded and stored on a server for access by users on the World Wide Web. The size of the product currently is not large, at 9.6Kb, but in theory creating a finite system that would include not only the additional functional nouns but also Mohawk verbs would be significantly larger. The program at present only has 30 noun roots but there are over 8,000 possible paths, most of which are valid words in Mohawk.

In referring back to Dougal Graham's (2007) Finite State Parser for Cayuga nouns, he was able to create two systems, one concrete, and one abstract, but he still never attempted to create a system that included Cayuga's functional nouns, that would have represented the structural complexity necessary to see if the system could have handled all Cayuga nouns and also given a preview for the feasibility of describing Cayuga verbs using finite state morphology. A few years after Graham's master's thesis, whether because a finite system would be too large or not for an e-dictionary and morphological parser, Carrie Dyck and Ranjeet Kumar do not utilize finite state morphology to create an e-dictionary (2012). This e-dictionary, while not completely inclusive, can chunk the Cayuga words and even provide suggestions for users while inputting a word. However, while comparing both products to any proposed tool for Mohawk is useful, they are ultimately not the same language and each have their unique characteristics. Both Cayuga and Mohawk being Iroquoian languages, they share many of the same traits and therefore what challenges face the development of NLP tools for Cayuga will also most likely be obstacles for Mohawk to overcome as well.

However, some differences even include community interest in the product. Dr. Dyck has been working with the Cayuga language for years and has many connections within the community and the community welcomed the dictionary projects, whereas currently there are few professional linguists working with Mohawk. While there are linguists that focus on the Iroquoian languages, there are few and connections with the Mohawk Nation do not seem to have had the same level of collaboration as with Dr. Dyck and the Cayuga Nation. Dr. Mariane Mithun of University of California Santa Barbara is one of the few whose current research involves not only the Iroquoian languages but also Mohawk. She is currently working toward publishing a complete grammar of the Mohawk Language in the next year, but this resource was not available at the time of the development of my system (Mithun, 2013). The grammar is slated to provide comparison of the regional variances in phonetics, syntax, and orthography among the different Mohawk communities. Perhaps this tool would help outline the hierarchy of adding multiple suffixes after the noun root which would make it possible to create that component in the morphological analyzer without having to reach out to Mohawk speakers and teachers from each community. Since each community has its differences, it

would be important to cross-reference and gather all the information about the morphotactics of this element. Additionally at the University of Buffalo, State University of New York is a professor Karin Michelson who has worked with Mohawk and the Iroquoian languages, and was involved in creating language resources for the Ontario Board of Education (Edu.gov.on.ca, 1999). Currently one of her prior students, a member of the Onondaga nation helped start up an Iroquois Linguistics Certificate Program through Syracuse University which is in its first run for the 2013-2014 academic year. While the community of researchers for Mohawk and Iroquoian languages is not large, there is active research and momentum is gaining for these languages.

7. Conclusion

The objective of this project was to create a morphological parser using finite state morphology to describe Mohawk formal nouns. The result is a program that contains a lexicon that contains the structure to describe most Mohawk formal nouns; however a rules program is yet to be written which will aid in the maintenance of structure for the Mohawk formal nouns, include spelling corrections and exceptions. The Mohawk language, like many of the world's languages is facing a likely future of extinction within the next century. However, it also shares in common the challenge that even many thriving languages are facing, and that is the development of NLP tools and the ease of use of the language in a digital platform. Currently focus for Mohawk is on language revitalization but also preservation, whereas some of the European languages are well-protected by their governments and educational systems, yet they still are fighting to have a digital presence. Working on NLP tools for Mohawk can hopefully directly benefit the Mohawk community itself, but due to its unique structure it also poses great challenges to computational linguists that can perhaps help push these researchers to develop more efficient and responsive NLP tools for other languages, after having thought about Mohawk. For instance, it is still not known whether or not the Iroquoian languages can be described using finite state morphology, even with the advanced functions that the Xerox software allows for features that the typical finite state system could not handle, such as long distance dependencies and reduplication. Neither work in Cayuga or this project with Mohawk is conclusive in answering that question, which would be an excellent proposal for future study. However, turning away from finite state morphology the current trend is in Machine Learning techniques, and Mohawk, unlike most of its fellow Native North American counterparts has rich documentation in Mohawk, stemming back from the 1600s. Applyinh a learning algorithm to parallel corpuses in English and Mohawk would be an interesting experiment to see if it would be possible and easier to create a morphological parser and in turn to create a simple machine translation tool or spell-checker function.

Educational material for the Mohawk language has increased in production in the last 10 years and on both sides of the US Canadian border in Ontario, New York State, and Quebec. Official resources are being published by the Departments of Education in each province and state to support the learning of the Iroquoian languages in a second language capacity. As for Maori in New Zealand, with Maori in schools and the ability to use Maori when navigating Microsoft Windows, digital tools need to be developed for Mohawk and the other Iroquoian languages for members of these communities to be able to function in their everyday lives in their language. For this reason programs and smartphone applications are being developed to allow the North American language communities to text in their own language, without have the iPhone autocorrect every word they attempt to spell out. Companies like Microsoft and Google are expanding their language offerings to include the under-resourced languages not only of North America but on every continent. In addition, non-profits like Mozilla are pursuing ways to motivate community members of languages exactly like Mohawk to get involved so their products such as the web browser Firefox might one day soon be available in Mohawk. It is also important to add that while it is wonderful the outside community has increased its awareness and interest in many of the world's under resourced and minority languages, far more important is that community members themselves recognize the value of their language and if not directly involved in the creation of a specific product that they support the development of these varying language tools. As a native speaker of English and a speaker of Spanish, Italian, and Russian, all of which are some of the most internationally well-known and studied languages of the world, the concept of a language as intellectual property that potentially means the words of a language are not free for use is a relatively difficult concept for me to grasp. Everyone can learn to speak English and I certainly do not control its use, but for many languages with smaller linguistic communities this is the case and work with Mohawk must be sensitive to this, and people working with any language community must respect the cultural aspects of the community and its language use. Microsoft found itself in a lawsuit when the Mapuche community along the Argentina and Chilean border region were angered by the language pack for Windows that Microsoft published in Mapudungun.

To these people, with the complex and tense history they have had with the colonial governments and then the governments of Argentina and Chile, the use of their language by non Mapuche and the use of it in a computer program was a violation of their rights, and they viewed it as an illegal infringement of their intellectual property. This is not to say that every community feels this way, but is a note of warning that when working with these languages, especially if a product is the end goal for distribution, that permission from the community and open communication and dialog is key. Many communities are all too familiar with outsiders and Europeans coming in and telling them what they need to help themselves.

Thus with this in mind, future research and development of NLP tools for Mohawk should be attempted with communication and collaboration with the community themselves to make sure that the products being developed are tools that the community wants and sees as useful. For the sake of linguistic research purely, it would be a great achievement to attempt to create an analyzer that could handle Mohawk verbs, or really attempting this with any Iroquoian language would be ground-breaking, since no one yet knows if it will be possible. Clearly an advanced understanding not only of the language itself and the rules, but also of computer programming and specifically of finite state morphology and the Xerox software is necessary before even beginning to attempt a project like this.

References

Aaanativearts.com 2013. *Native American Tribes by Language*. [online] Available at: http://www.aaanativearts.com/native-american-tribes-by-language.html#axzz2hnx7Dsmv [Accessed: 15 Oct 2013].

An, Kumar, M., Dhanalakshmi, V., Rekha, R., Soman, K. and Rajendran, S. 2010. A Novel Data Driven Algorithm for Tamil Morphological Generator. *International Journal of Computer Applications*, pp. 52--56.

Antworth, E. 1992. Glossing Text with the PC-KIMMO Morphological Parser. *Computers and the Humanities*, 26 (5/6), p.389-398.

Antony, P. and Soman, K.. Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey.

Baker, M. 1996. *The polysynthesis parameter*. New York [u.a.]: Oxford Univ. Press.

Baker, M. 2002. *The Atoms of Language*. New York: Basic Books.

Beesley, K. and Karttunen, L. 2003. *Finite state morphology*. Stanford, Calif.: CSLI Publications.

Bonvillain, N. 1973. A Grammar of Akwesasne Mohawk. Ottawa, Canada: National Museums of Canada, Mercury Series.

Bogaert, H., Wassenaer, N. and Megapolensis, J. 1999. *Early vocabularies of Mohawk.* Southampton, Pa.: Evolution Pub..

Bosch, S., Jones, J., Pretorius, L. and Anderson, W. 2007. Computational Morphological Analysers and Machine-Readable Lexicons for South African Bantu Languages. *Localisation Focus*, p. 22.

Bruyas, J. 1862. *Radical words of the Mohawk language with their derivatives.* New York: Cramoisy Press.

Chafe, W. 2007. *Handbook of the Seneca language*. Toronto: Global Language Press.

Chafe, W. 2012. Are adjectives universal? The case of Northern Iroquoian. *Linguistic Typology*, 16 p.1-39.

Chafe, W. and Foster, M. 1981. Prehistoric divergences and recontacts between Cayuga, Seneca, and the other Northern Iroquoian Languages. *International Journal of American Linguistics*, 47 (2), pp. 121--142.

Charles River Editors 2012. *Native American Tribes: The History and Culture of the Iroquois Confederacy*. [e-book] Amazon Digital Services, Inc..

Deering, N. and Harries-Delisle, H. 2007. *Mohawk A Teaching Grammar*. Kahnawa:ka, Canada: Kanien'keha:ka Onkwawen:na Raotitiohkwa Language and Cultural Center.

Dyck, C. 2009. Defining the word in Cayuga (Iroquoian). *International Journal of American Linguistics*, 75 (4), pp. 571--605.

Dyck, C. and Kumar, R. 2012. A grammar-driven bilingual digital dictionary for Cayuga (Iroquoian).*Dictionaries: Journal of the Dictionary Society of North America*, 33 (1), pp. 179--204.

Edu.gov.on.ca. 1999. *Native Languages*. [online] Available at: http://www.edu.gov.on.ca/eng/curriculum/secondary/nativelang.html [Accessed: 19 May 2013].

Ethnologue.com. 2013. *Welcome | Ethnologue*. [online] Available at: http://www.ethnologue.com/ [Accessed: 19 May 2013].

Gate.ac.uk. 2013. *GATE.ac.uk - overview.html*. [online] Available at: http://gate.ac.uk/overview.html [Accessed: 14 Oct 2013].

Graham, D. 2007. *Finite State Parsing of Cayuga Morphology*. Master of Arts. Memorial University.

Haribhakta, Y., Kalamkar, S. and Kulkarni, P. 2012. Feature annotation for text categorization. pp. 308--313.

Hasler, L. 2013. *Cayuga: Our Oral Legacy - Home*. [online] Available at: http://cayugalanguage.ca/ [Accessed: 15 Oct 2013].

Hickerson, H., Turner, G. and Hickerson, N. 1952. Testing procedures for estimating transfer of information among Iroquois dialects and languages. *International journal of American linguistics*, 18 (1), pp. 1--8.

Hill, H., Hess, W., Wilkes, J. 1835. *The Acts of the Apostles, in the Mohawk Language.* New York: Young Men's Bible Society of New York.

Hinrichs, E. 2005. Finite-state parsing of German. *Inquiries into Words, Constraints and Contexts*, p. 35.

Homola, P. 2011. Parsing a Polysynthetic Language.. pp. 562--567.

Homola, P.. A Machine Translation Toolchain for Polysynthetic Languages.

Inuktitutcomputing.ca. 2000. *Inuktitut Morphological Analyzer*. [online] Available at: http://www.inuktitutcomputing.ca/Uqailaut/en/IMA.html [Accessed: 19 May 2013].

Iroquoismuseum.org. 2013. *Iroquois Indian Museum*. [online] Available at: http://www.iroquoismuseum.org/ [Accessed: 19 May 2013].

Idibon. 2013. *Home - Idibon*. [online] Available at: http://idibon.com/ [Accessed: 14 Oct 2013].

Jimerson, R. 2013. *Interview on Seneca Language Project*. Interviewed by Alicia Alexandra Assini [in person] Dublin, Ireland, April 15, 2013.

Jones, J. 2013. *Star Wars Gets Dubbed into Navajo: a Fun Way to Preserve and Teach a Fading Language*. [online] Available at: http://www.openculture.com/2013/07/star-wars-gets-dubbed-into-navajo.html [Accessed: 15 Oct 2013].

Kanienkehaka.com. 2013. *The Mohawk Language Standardisation Project > Ministry of Education / Ministry of Training, Colleges and Universities*. [online] Available at: http://www.kanienkehaka.com/msp/msp.htm [Accessed: 15 Oct 2013].

Karlsson, F. and Koskenniemi, K. 1985. A Process Model of Morphology and Lexicon. *Folia Linguistica*, 1 (1/2), p.207-231.

Karttunen, L. 1983. KIMMO: a general morphological processor. 22 pp. 163--186.

Khan, N. and Khan, M. 2006. Developing a computational grammar for Bengali using the HPSG formalism.*Center for Research on Bangla Language Processing, BRAC University*.

Koskenniemi, K. 1983. Two-Level Model for Morphological Analysis.. 83 pp. 683--685.

Long, S. 2012. Ranking Native American language health. *techna verba scripta*, [blog] 26th Sept 2012, Available at: http://technaverbascripta.wordpress.com/2012/09/26/ranking-native-american-language-health/ [Accessed: 18th May 2013].

Lonsdale, D. and Matsushita, H. n.d. Annotating and exploring Lushootseed morphosyntax. Available at: http://linguistics.byu.edu/faculty/lonsdaled/lutportal1.pdf [Accessed: 16th May 2013].

Maracle, D. 1993. *Kanyen'keha Tewatati Let's Speak Mohawk*. Guilford, Conn: Audio-Forum.

Maracle, D. 1992. *One Thousand Useful Mohawk Words*. Madison, Connecticut: Audio Forum.

Megerdoomian, K. n.d. *Finite State Morphology: A Tutorial*. [e-book] http://www.zoorna.org/handouts/FSM.pdf [Accessed: 17th May, 2013].

Michelson, K. 1981. A Philological Investigation into Seventeenth-Century Mohawk. *International Journal of American Linguistics*, 47 (2), pp. 91--102.

Mithun, M. 2013. *Interview about Mohawk Language resources.* Interviewed by Alicia Alexandra Assini [via Skype] Buenos Aires, Argentina, July 26, 2013.

Mithun, M. 1996. Grammatical Sketches: The Mohawk Language. In: Maurais, J. eds. 1996. *Quebec's Aboriginal Languages.* Toronto: Multilingual Matters Ltd, pp. 159-173.

Mithun, M. and Corbett, G. 1999. The effect of noun incorporation on argument structure. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pp. 49--72.

Mithun, M. 2005. Grammar and the Community. *Studies in Language*, 30 (2), p.281-306.

Muller, K., Mika, S., Ratsch, G., Tsuda, K. and Scholkopf, B. 2001. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12 (2), pp. 181--201.

Name, Y. 2013. *Home*. [online] Available at: http://www.ndnlanguage.com/index.html [Accessed: 14 Oct 2013].

News.ualberta.ca. 2013. *E-books show kids the colour of Cree language - News & Events - University of Alberta*. [online] Available at: http://news.ualberta.ca/newsarticles/2013/january/ebooksshowkidscolourofcree [Accessed: 14 Oct 2013].

Ogokilearning.com. 2013. *About Ogoki Learning Systems | Ogoki Learning Systems*. [online] Available at: http://www.ogokilearning.com/about-ogoki-learning-systems-learning-president-bio/ [Accessed: 14 Oct 2013].

Olan, M. 2013. *Interview on Mohawk Language Revitalization Efforts*. Interviewed by Alicia Alexandra Assini [in person] Dublin, Ireland, 04/10/2013.

Ptaszynski, M. and Kazuki, M., et al. 2013. Untitled paper, paper presented at untitled conference, NLP for Endangered Languages: Morphology Analysis, Translation Support and Shallow Parsing of Ainu Language. Association for Natural Language Processing, p.418-421.

Rambow, O., Bangalore, S., Butt, T., Nasr, A. and Sproat, R. 2002. Creating a finite-state parser with application semantics. pp. 1--5.

Seiss, M. and Nordlinger, R. 2011. An electronic dictionary and translation system for Murrinh-Patha.

Talkmohawk.com. 2013. *Press Release: Mohawk Language Mobile App | Talk Mohawk .com*. [online] Available at: http://www.talkmohawk.com/blog/press-release/press-release-mohawk-language-mobile-app/ [Accessed: 14 Oct 2013].

Technaverbascripta.wordpress.com. 2013. *Ranking Native American language health*. [online] Available at: http://technaverbascripta.wordpress.com/2012/09/26/ranking-native-american-language-health/ [Accessed: 15 Oct 2013].

Www12.statcan.gc.ca. 2013. *Aboriginal languages in Canada*. [online] Available at: http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011003_3-eng.cfm [Accessed: 15 Oct 2013].

YouTube. 2013. *LRC X8 - Reinhard Schäler - The Rocky Localisation Picture Show*. [online] Available at: http://www.youtube.com/watch?v=7akCTvsOdic [Accessed: 4 Nov 2013].

Appendix

A. *Lexc* program code for Morphological Analyser

```
Multichar_Symbols +nomPF +nomSF +possPF1p +possPF1p2P +possPF1pP +possPF2pS
+possPF2p2P +possPF2pP +possPF3pmS +possPF3pfS +possPF3pnS +possPF3pmP +possPF3pfS
+locSF1 +PLAn +PLInam

LEXICON      Root
                    basicNouns      ;
                    defectiveNouns ;

LEXICON                             defectiveNouns
kitkit: chicken                             NSFs ;
kwiskwis: pig                               NSFs ;
tawistawis: plover                          NSFs ;
takois: cat                                 NSFs ;

LEXICON                                     basicNouns
                                            basicPossPF;
                                            basicUnpossPF;

LEXICON                                     basicUnpossPF
ka: +nomPF                                  basicKaRoots;
o: +nomPF                                   basicORoots;
a: +nomPF                                   basicARoots;
aw:  +nomPF                         basicERoots;
                                            basicARoots;
                                            basicERoots;

LEXICON                             basicPossPF
ake: +possPF1p                              basicKaRoots ;
ak: +possPF1p                               basicORoots ;
akw: +possPF1p                              basicARoots ;
sa:  +possPF2pS                             basicAllRoots ;
rao:  +possPF3pmS                           basicAllRoots ;
ak(a)o: +possPF3pfS                         basicAllRoots ;
ao: +possPF3pnS                             basicAllRoots ;
onkeni: +possPF1p2P                         basicKaRoots ;
                                            basicORoots ;
onkwa: +possPF1pP                           basicARoots ;
                                            basicERoots ;
seni: +possPF2p2P                           basicAllRoots ;
sewa: +possPF2pP                            basicAllRoots ;
raoti: +possPF3pmP                          basicKaRoots ;
                                            basicORoots ;
```

```
raon: +possPF3pmP                basicARoots ;
                                 basicERoots ;
aoti: +possPF3pfP                basicKaRoots ;
                                 basicORoots ;
aon: +possPF3pfP                 basicARoots ;
                                 basicERoots ;

LEXICON                          basicAllRoots
                                 basicARoots ;
                                 basicERoots ;
                                 basicKaRoots ;
                                 basicORoots ;

LEXICON                          basicKaRoots
nat: town                        NSFs ;
nonhs: house                     NSFs;
nakt: bed                        NSFs ;
tshe: bottle                     NSFs ;
honwey: boat                     NSFs ;
ront: tree                       NSFs ;


LEXICON                          basicORoots
hont:grass                       NSFs;
tsitsy(a): flower                NSFs;
ryent: habit                     NSFs;
weyenn: ability                  NSFs;
nikonr: spirit                   NSFs;
wenn: word                       NSFs;
nhwentsy: earth                  NSFs;
nhwentsya: ground                NSFs;
wir: corn                        NSFs;
hsin: leg                        NSFs;
konhs: face                      NSFs;

LEXICON                          basicARoots
aht: footwear                    NSFs;
ather: basket                    NSFs;
atoken: axe                      NSFs;
ashar: knife                     NSFs;
arhy: hook                       NSFs;
ahsir: blanket                   NSFs;

LEXICON                          basicERoots
erhar: dog                       NSFs;
ehsa: blackash                   NSFs;
eryahs: heart                    NSFs;

LEXICON                          NSFs
                                 basicSF;
                                 locSF;
                                 attSF;

LEXICON                          basicSF
a: +nomSF                        #;
```

```
e: +nomSF                       #;
                                #;

LEXICON                         locSF
                                locfinalSF;
atsher: +locSF1                 locfinalSF;
asher: +locSF1                  locfinalSF;
ahkw: +locSF1                   locfinalSF;

LEXICON                         locfinalSF
ake: on                         # ;
hne: on                         # ;
akon: in                        # ;
okon: under                     # ;
akta: near                      # ;

LEXICON                         attSF
okon: +PLAn                     # ;
okonha: +PLInan                 # ;
iyo: good                       # ;
aksen: bad                      # ;
akayon: old                     # ;
ase: new                        # ;
owanen: big                     # ;
kowa: big                       # ;


END 6656
```