Departament de Traducció i Ciències del Llenguatge
Universitat Pompeu Fabra

Programa de doctorado:
Traducció i Ciències del Llenguatge

Plan de Investigación Doctoral

Dungupeyem: Analizador y Generador Morfológico a través de Transductores de Estados Finitos. Y otras Herramientas Lingüístico-Computacionales para el Mapudungun

Plan de tesis doctoral dirigido por el Dr. Toni Badia
Andrés Chandía

Barcelona, 2013.

Abstract

The work proposed by this research project is to build computational linguistic tools capable of processing texts written in Mapudungun.

Computational applications are becoming increasingly sophisticated and their use is being extended to a wider variety of areas. However, the less represented languages remain behind in terms of the NLP tools available to process them. Fortunately, there are already a few initiatives to create computational resources to enable minority languages to become part of the digital world and there are some signs that the interest among computational linguists in this kind of enterprise is growing.

The main goal of this project is to start creating the necessary tools that would bring the Mapudungun language into the growing group of languages with linguistically annotated electronic corpora. Our main target is the implementation of a morphological analyzer capable of processing the linguistic input and generating idiomatic well formed units of this language. In order to achieve the desired results, we will have to effectively deal with some problems such as the lack of a standardized orthographic system for Mapudungun. We believe, however, that the techniques of finite state transducers (FST) are appropriate to deal with some of the most important problems posed by the Mapudungun morphophonology. FST techniques will be used not only to implement the morphological analyzer and generator (MAG), but also the tokenizer, orthographic normalizer, code normalizer, morphological guesser, spell checker and verifier, part of speech disambiguator and tagger.

A second goal it is to figure out the pending or confusing subjects of the Mapudungun description, and also investigate the techniques for unsupervised morphological analysis and generation.

Resum

El treball proposat per aquest projecte d'investigació és construir eines lingüisticocomputacionals que permetin el tractament de texts escrits en Mapudungun.

Les aplicacions computacionals són cada vegada més sofisticades i inclouen cada dia més àmbits, tanmateix, en el camp de les eines lingüístiques computacionals sempre van quedant enrere les llengües amb menys presència, afortunadament existeix un incipient corrent de lingüistes computacionals que està donant cabuda a les llengües minoritàries al món digital.

El propòsit central d'aquest projecte és crear les eines necessàries per apropar el Mapudungun a aquest creixent grup de llengües amb còrpora lingüísticament anotat. El nostre objectiu principal és la implementació d'un analitzador morfològic capaç de processar els inputs lingüístics, i de generar unitats idiomàtiques ben formades d'aquesta llengua. Per assolir aquests resultats haurem d'enfrontar-nos a alguns problemes com la falta d'un sistema ortogràfic estandarditzat del Mapudungun, per exemple. Tanmateix, creiem que les tècniques de transductors d'estats finits (FST) són les apropiades per enfrontar la casuística de la morfofonología del Mapudungun. Utilitzarem els FST no només per implementar el analitzador i generador morfològics (MAG), sinó també tokenitzadors, normalitzadors ortogràfics, normalitzadors de codi, guessers morfològics, verificadors i correctors ortogràfics, desambiguadors i etiquetadors de les parts de l'oració.

Un segon objectiu és desentranyar els temes pendents o confusos de la descripció del Mapudungun, com també investigar les tècniques d'anàlisi i generació morfològiques no supervisades.

Resumen

El trabajo propuesto por este proyecto de investigación es construir herramientas lingüístico-computacionales que permitan el tratamiento de textos escritos en Mapudungun.

Las aplicaciones computacionales son cada vez más sofisticadas y abarcan cada día más ámbitos, sin embargo, en el campo de las herramientas lingüísticas computacionales siempre van quedando atrás las lenguas con menos presencia, afortunadamente existe una incipiente corriente de lingüistas computacionales que está dando cabida a las lenguas minoritarias en el mundo digital.

El propósito central de este proyecto es crear las herramientas necesarias para acercar el Mapudungun a este creciente grupo de lenguas con córpora lingüísticamente anotado. Nuestro objetivo principal es la implementación de un analizador morfológico capaz de procesar los inputs lingüísticos, y de generar unidades idiomáticas bien formadas de esta lengua. Para alcanzar estos resultados tendremos que enfrentarnos a algunos problemas como la falta de un sistema ortográfico estandarizado del Mapudungun, por ejemplo. Sin embargo, creemos que las técnicas de transductores de estados finitos (FST) son las apropiadas para enfrentar la casuística de la morfofonología del Mapudungun. Utilizaremos los FST no sólo para implementar el analizador y generador morfológicos (MAG), sino también tokenizadores, normalizadores ortográficos, normalizadores de código, guessers morfológicos, verificadores y correctores ortográficos, desambiguadores y etiquetadores de las partes de la oración.

Un segundo objetivo es desentrañar los temas pendientes o confusos de la descripción del Mapudungun, como también investigar las técnicas de análisis y generación morfológicas no supervisadas.

Índice de contenido

1. Presentación del tema de investigación

Nuestra propuesta se enmarca en el ámbito de la lingüística computacional y tiene relación con el desarrollo de herramientas informáticas para lenguas minoritarias, la investigación sobre aplicaciones de ejecución no supervisada, y la definición de ciertas reglas de combinación entre raíces y sufijos en el Mapudungun, el idioma de los mapuche, un pueblo aborigen de América del Sur que hoy habita territorios ubicados al sur de Chile y Argentina, aunque la migración a las ciudades ha restaurado su presencia en las zonas norteñas de ambos países; de todas maneras, los mapuche son más numerosos en el lado oeste de la Cordillera de los Andes, en Chile. La cantidad de habitantes mapuche se estima en cerca de setecientas mil personas (los datos varían según los censos), de los cuales sólo un 30%, aproximadamente, son hablantes, con diferentes grados de competencia, de Mapudungun.

El Mapudungun es una lengua de tipo polisintético aglutinante, es decir, tiene una estructura interna compleja, compuesta por una serie de morfemas que se combinan según pautas muy acotadas y, a su vez, estos morfemas son claramente segmentables. Otra característica importante de esta lengua es la incorporación de palabras completas, incluso series de palabras, en la estructura del verbo (Salas, 2006: 56).

Afortunadamente existen diversos estudios lingüísticos sobre el Mapudungun, algunos muy extensos, por lo que no se camina tan a ciegas al momento de desarrollar instrumentos informáticos que permitan la exploración de esta lengua. Sin embargo, todavía hay mucho por documentar sobre la lengua de los mapuche, situación que hemos comprobado al intentar definir por ejemplo, qué tipo de raíces nominales y adjetivales son las que combinan con uno u otro tipo de sufijos verbalizadores, los que posibilitan que dichas raíces puedan generar un predicado verbal; esta situación se repite en otros ámbitos de la lengua que

no han sido estudiados pormenorizadamente. Creemos que el desarrollo de herramientas lingüístico-computacionales para el Mapudungun pondrá al servicio de los lingüistas, y de quien quiera utilizarlas, medios valiosos para describir aquellos fenómenos que aún siguen sin ser comprendidos o estudiados.

Dentro de la explosión actual en la cantidad de información y los medios para acceder a ella, una buena parte del mundo ha quedado atrás porque la información se ha generado en las lenguas predominantes, sobre todo en inglés. Los miles de idiomas utilizados por las minorías tienen una representatividad mínima cuando no, nula. Concentrar esfuerzos en una lengua minoritaria como el Mapudungun, tenemos la esperanza, rendirá frutos aplicables a otras lenguas minoritarias de la misma tipología, en general lenguas de la zona de América del Sur, que son incluso más desconocidas que la que nosotros tratamos.

El proyecto que proponemos es, según lo vemos, sólo el principio de un desarrollo continuo de herramientas computacionales aplicadas al estudio de las lenguas. Nosotros comenzaremos con un MAG (Morphological Analizer and Generator) basado en FST. Este sistema ya hemos comenzado a ponerlo en práctica, y pensamos que llevamos un buen camino recorrido, hemos logrado incluir el análisis y generación del predicado verbal mapuche, el eje y parte más extensa y complicada de esta lengua, al menos en sus formas más elementales. Nos falta concluir y afinar el trabajo hecho sobre las formas verbales, incluir, por ejemplo, la movilidad de algunos sufijos, la oración subordinada o las frases nominalizadas, y extenderlo a las otras partes de la lengua.

El trabajo que hemos estado haciendo se basa en un compilador de FST desarrollado en el Centro de investigación lingüístico-computacional de la compañía Xerox, por lo tanto, el software que genere este compilador será propietario y de código protegido, lo que quiere decir que es de distribución comercial previo pago de las licencias pertinentes. Debido a que nuestro objeto de estudio es una legua minoritaria, además, perteneciente a un pueblo empobrecido y discriminado durante siglos, nos parece que la mejor manera de aportar a la difusión, conservación y estudio del Mapudungun es generar herramientas de libre

distribución y código abierto. En este sentido, hoy en día existe una amplia gama de posibilidades, compiladores como FOMA[1] o HFST[2] que básicamente funcionan como el XFST[3] de Xerox, pero con licencia de libre distribución. U otras opciones desarrolladas con lenguajes de programación ideados para el fin que nosotros buscamos, nos estamos refiriendo específicamente a Python[4]. Aunque no somos expertos en este lenguaje de programación, sabemos que existen y conocemos una serie de herramientas desarrolladas con Python para fines específicamente lingüísticos, de hecho, es de conocimiento de todo lingüista que se dedica al análisis computacional, la existencia del proyecto NLTK[5] (Natural Language Toolkit), una plataforma para escribir programas con el lenguaje Python dedicados al tratamiento del lenguaje natural, a través del trabajo sobre córpora, categorización de texto, análisis de estructuras lingüísticas, y muchas otras implementaciones.

Además, durante el desarrollo de nuestro trabajo previo, nos hemos topado con una iniciativa del grupo de investigación "Human Language Technology and the Democratization of Information[6]" de la Universidad de Indiana en los Estados Unidos de Norteamérica. El proyecto se llama "L^3" ("Learning Lots of Languages[7]"), y define su horizonte en estos términos: El L^3 tiene el objetivo, a largo plazo, de desarrollar un sistema para traducir desde y hacia muchas lenguas débilmente representadas y pertenecientes al Sur global. Y, menos ambiciosamente, de crear herramientas para la extracción de información y para el aprendizaje asistido por ordenador de estas lenguas.

Idealmente nuestro objetivo es el mismo, pero desde una posición realista, sabemos que es un trabajo a muy largo plazo, por ello hemos acotado nuestras perspectivas. Como ya habíamos mencionado, la primera tarea será afinar el sistema que ya tenemos funcionando, luego incluiremos las variantes de las for-

1 https://code.google.com/p/foma/
2 http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/
3 http://www.stanford.edu/~laurik/fsmbook/home.html
4 http://www.python.org/
5 http://nltk.org/
6 http://www.cs.indiana.edu/~gasser/Research/hltdi.html
7 http://www.cs.indiana.edu/~gasser/Research/projects.html

mas verbales. El paso siguiente será trabajar en el desarrollo de reglas para el tratamiento de las demás partes de la lengua mapuche, sustantivos, adjetivos, adverbios, etc. Una vez que consideremos maduro el sistema de análisis y generación morfológicos para el Mapudungun nos volcaremos a la tarea de implementarlo dentro del marco del código abierto, es decir, traducir nuestro sistema al entorno de libre distribución, idealmente, y dependiendo de la complejidad, sería el ambiente Python desarrollado por el grupo de la Universidad de Indiana. Para ello seguiríamos las directrices del Doctor Michael Gasser[8], uno de los principales desarrolladores de herramientas lingüístico-computacionales con que cuenta este grupo.

Habiendo completado el proceso del analizador morfológico, nos centraremos en el desarrollo de otras herramientas que se pueden derivar de este trabajo, como por ejemplo tokenizadores, normalizadores ortográficos, normalizadores de código, guessers morfológicos, verificadores y correctores ortográficos, desambiguadores y etiquetadores de partes de la oración.

Para afianzar la vertiente investigadora que nos mueve, intentaremos poner en práctica otro mecanismo de análisis morfológico que nos permitirá tratar e incorporar formas no listadas en nuestro sistema, nos referimos a los sistemas basados en cálculos matemáticos de autogestión, o no supervisados. Un sistema entrenado para realizar tareas controladas por patrones de aparición recogidos de cantidades masivas de datos. Los datos relevantes son textos, en jerga lingüística, corpus sin anotar, que son introducidos en en las aplicaciones que se encargan, mediante algorit,mos, de reconocer y discriminar raíces léxicas de prefijos y sufijos, por ejemplo. Es imprescindible prestar atención a estas técnicas y dedicar algún tiempo a entenderlas, probarlas y evaluarlas. Centrarnos en la investigación de estos métodos es necesario si queremos robustecer nuestro sistema sin invertir excesivo tiempo. En lo referente a la descripción del Mapudungun, nos dedicaremos a investigar cómo deberíamos clasificar las raíces nominales y adjetivales que combinan con uno u otro tipo de sufijo verbalizador para generar frases predicativas o frases verbales nominalizadas.

8 http://www.cs.indiana.edu/~gasser/

1.1. Problema de investigación u objeto de análisis

Como ya hemos adelantado, nuestro objeto de estudio es el Mapudungun, más estrictamente sus composición y fenomenología morfofonológicas y, hasta cierto punto, sintácticas. Al ser una lengua aglutinante y polisintética, como la hemos descrito en la sección anterior, la sintaxis y la morfología forman un tejido imbricado inseparable, es por ello que hemos hablado de incluir, por ejemplo, las oraciones subordinadas, porque éstas se forman de la misma manera que las oraciones simples, pero combinando diferentes morfemas.

Nuestros objetivos son implementar el MAG, investigar herramientas cuyos procesos no necesiten supervisión para integrarlas con el MAG y que ejecuten tareas adicionales y complementarias en el procesamiento del lenguaje natural; intentar definir los tipos de raíces nominales y adjetivales que combinan con los diferentes tipos de sufijos verbalizadores. Para el primero de ellos debemos seguir alguna directriz, una descripción morfofonológica lo suficientemente clara y concisa como para volcarla a un sistema de FST. La investigadora Ineke Smeets publicó el año 2008 su tesis doctoral del año 1989, una descripción gramatical del Mapudungun muy pormenorizada y clara. *A Grammar of Mapuche* (Smeets, 2008) es la piedra angular de nuestro proyecto, pues intentamos implementar en el sistema de FST la descripción de la gramática del Mapudungun hecha por la autora. Las partes del libro que cimientan nuestro trabajo son la *V: Morphology and morphosyntax of the verb* y la *III: Morphology and morphosyntax of the noun*.

En la parte *V* encontramos una descripción de los distintos sufijos que siguen a la raíz verbal mapuche, estos corresponden a diferentes tipos, empezando por los más cercanos a la raíz: modificadores de valencia, aspectuales, modificadores semánticos, de valor de verdad, y flexión. La autora ubica los afijos en 36 slots, algunos de los afijos tienen cierta movilidad, y muchos de ellos tienen interdependencias, haciendo que algunos estén obligatoriamente presentes o ausentes según la presencia o ausencia de otros de los afijos.

El predicado verbal mapuche se expresa al estar una forma finita del verbo seguida por los sufijos, de los cuales es obligatorio el de sujeto en el slot 3, y este obliga a su vez la presencia del marcador de número en el slot 2. Los marcadores modales, también obligatorios, ocupan el slot 4, indicativo, condicional[9] e imperativo. Por su parte, una frase subordinada se expresa mediante un verbo no finito, que no utiliza el marcador de sujeto, pero que contiene obligatoriamente un morfema flexivo de nominalización en el slot 4. Una forma verbal es seguida de, al menos, un sufijo y en raras ocasiones sobrepasa los diez, la autora comenta que la palabra que ha encontrado con más sufijos, contenía trece; las frases subordinadas suelen tomar menos sufijos que las predicativas.

En esta parte de libro también se tratan las raíces verbales y los auxiliares (cap. 25)[10]. En el capítulo siguiente trata la morfología verbal y la posición de cada morfema en los 36 slots (cap. 26). Ubica entre los capítulos 27 y 31 el tratamiento de sufijos de escasa aparición, nominalizadores derivativos, composición y, finalmente, verbos deícticos y defectivos.

En la parte III, que versa sobre la morfología y morfosintaxis del sustantivo, especialmente importantes son los capítulos 18 a 21; en ellos trata la sufijación con respecto a las raíces nominales, diferenciando los sufijos que no cambian la clase nominal de la raíz de los que si la cambian (cap. 18). Luego trata otros fenómenos morfosintácticos como la composición (cap. 19), la reduplicación (cap. 20) y, por último, la verbalización, o mejor dicho los morfemas verbalizadores que pueden convertir en verbo a adjetivos, adverbios, numerales y sustantivos (cap. 21).

Otros temas que nos son de extrema utilidad aparecen en los capítulos 10 al 17, también de la parte III: adjetivos, adverbios, pronombres demostrativos y anafóricos, pronombres personales, pronombres posesivos y pronombres interrogativos. La morfología y la fonología están tratadas en la parte II del libro, el capítulo 5 trata la estructura fonémica de las raíces, sufijos y palabras. El capítulo

9 Lo que Smeets califica de condicional, Zúñiga lo identifica como subjuntivo (Zúñiga, 2006).

10 El libro de Smeets se divide en 9 partes, y cada parte contiene sus capítulos que varían en número, el total de capítulos es de 35.

6 trata la distribución fonémica en raíces, sufijos y en las fronteras entre morfemas. Luego, el capítulo 8 describe la morfofonología, las variaciones consonánticas y vocálicas, las secuencias vocálicas y la inserción de fonemas.

Como hemos especificado, nuestro primer trabajo será plasmar la morfofonología del Mapudungun en un sistema de FST capaz de ejecutar procesos de análisis y generación basados en las reglas que iremos estipulando para este fin. Para esta tarea es fundamental seguir una metodología de trabajo que comprenda los mecanismos y reglas inherentes a los procesos descritos, es preciso utilizar cierto lenguaje de codificación. Por el reconocido prestigio que tiene, y por la documentación de que se dispone hemos decidido poner en práctica las directrices publicadas por Beesley y Karttunen, utilizando principalmente su libro *Finite State Morphology*, del año 2003, en donde se explica el sistema de FST desarrollado en la empresa Xerox (Beesly & Karttunen, 2003).

El XFST[11] es un conjunto de herramientas integradas para crear redes de estados finitos (FSN), es decir, sucesiones de transductores, que ejecutan los procesos de análisis y generación. Mediante la interfaz interactiva de Xerox se tiene acceso a los algoritmos de cálculo. El XFST también cuenta con un compilador de metalenguaje escrito a través de expresiones regulares (RE) que soporta las reglas de reemplazo, y que ayudan a generar procesos más fluidos y de codificación más intuitiva.

En el campo de los FST hay dos aspectos que concentran los esfuerzos al intentar expresar mediante sus algoritmos las reglas morfofonológicas de una lengua: la morfosintaxis o morfotáctica, y las alternancias fonológicas y ortográficas. La morfotáctica aborda la composición de las palabras por unidades más pequeñas, los morfemas, que están condicionados a aparecer en cierto orden o combinación. Las alternancias fonológicas y ortográficas, como se puede suponer, son cambios en la pronunciación o escritura de las palabras, debidos en gran parte al entorno en que se encuentran las partes afectadas. No sólo son cambios de una vocal por otra, por ejemplo, a veces desaparecen ciertos sonidos y otras,

11 http://www.stanford.edu/~laurik/fsmbook/home.html

se incorporan y, generalmente, estos fenómenos se ven reflejados en la escritura.

Ambos aspectos pueden ser tratados utilizando los sistemas de estados finitos, las combinaciones correctas de morfemas pueden ser expresadas mediante FSN, y las reglas que determinan la forma de cada morfema según contexto, pueden ser implementadas en los FST.

1.2. Estado de la cuestión

1.2.1. Estudio de la morfología del Mapudungun

En esta parte resumiremos la información expuesta en *La lengua mapuche (mapudungu(n)) hablada en Chile: sus principales rasgos estructurales* (Oyarzo, 2008: 8-12). Presentaremos los estudios de la lengua mapuche en 4 períodos, Colonial, Moderno, Posmoderno[12] y Contemporáneo[13].

Los primeros estudios gramaticales sobre la lengua mapuche fueron realizados por sacerdotes jesuitas que llegaron a Chile con propósitos de evangelización, este es el período colonial. *Arte y Gramática General de la Lengua que corre en todo el Reyno de Chile, con un Vocabulario y Confessonario,* del jesuita español Luis de Valdivia, Lima 1606. *Arte de la Lengua General del Reyno de Chile, con un diálogo chileno-hispano muy curioso,* del jesuita catalán Andreu Febrés, Lima 1765. *Chilidúgú Sive Res Chilenses vel Descriptio Status, tum civilis, cum moralis Regni populique chilensis,* del jesuita alemán Bernhard Havestadt, Alemania 1777. Todas tienen en común el carácter pedagógico (Salas, 1992: 477), dirigido hacia los sacerdotes que debían aprender la lengua vernácula para evangelizar a los aborígenes.

12 Utilizamos este término tanto porque es un período posterior al que Oyarzo calificó de moderno, y porque más o menos coincide con la época del posmodernismo.

13 Oyarzo describe 3 períodos: Estudios coloniales, Estudios modernos y Estudios contemporáneos, estos últimos los renombramos como postmodernos, para definir como contemporáneos a los estudios hechos en la década del 2000, que Oyarzo no incorpora.

En el período moderno destacó el profesor Rudolf Lenz (1863-1938), quien comenzó estudiando la influencia de las lenguas aborígenes sobre el castellano, momento en que desarrolló su interés por el Mapudungun. Revisó las gramáticas del período colonial y concluyó que el material lingüístico presentado no era auténtico y que la teoría gramatical no era adecuada, pues correspondía al modelo de las lenguas clásicas europeas que, aplicado al mapuche producía una visión distorsionada de su gramática. Lenz esbozó la constitución de un córpora en Mapudungun para analizarlo de acuerdo con las técnicas y conocimientos de la lingüística moderna. Las investigaciones de Lenz sobre lengua y cultura mapuche se reunieron bajo el título *Estudios Araucanos* (1895-1897). La comparación entre el material lingüístico de Lenz y el presentado por las fuentes coloniales permitió reconocer el alto grado de estabilidad estructural del Mapudungun.

Otro destacado es el monje capuchino alemán fray Félix José de Augusta quien publicó *Gramática Araucana* en1903, *Lecturas Araucanas* en 1910 y el *Diccionario Araucano-Español. Español-Araucano* en 1916. A diferencia de las Artes jesuitas, éstas presentan características de valor científico, pues el material lingüístico fue obtenido de hablantes nativos y se procesó de acuerdo a métodos modernos de la lingüística. Salas destaca la importancia del *Diccionario Araucano-Español - Español-Araucano* al que califica como una "obra impresionante", tanto por la cantidad de entradas como por la calidad lexicográfica (Salas, 1992: 487).

Un tercer autor del período moderno es el sacerdote capuchino fray Ernst Wilhelm von Mösbach. Tres obras importantes de este autor son: *Vida y costumbres de los indígenas araucanos en la segunda mitad del siglo XIX* de 1930; *Voz de Arauco. Explicación de los nombres indígenas de Chile* de 1944, e *Idioma mapuche, dilucidado y descrito con aprovechamiento de la Gramática Araucana del Padre Félix de Augusta* de 1962. El último es una gramática con propósito descriptivo y no pedagógico, según su propio autor.

En el período que hemos llamado posmoderno hay una tendencia a tratar aspectos lingüísticos específicos. Y se utilizan teorías de distintas corrientes lingüísticas. Encontramos a autores chilenos y extranjeros publicados en revistas

académicas especializadas. Adalberto Salas, investigó principalmente la fonología y la morfología verbal; representativas de estos ámbitos son: *Mapuche-español: análisis fonológico contrastivo* de 1978 y *Paradigma mínimo de las formas verbales del mapudungu, lengua de los mapuches o araucanos del centro sur de Chile* de 1980-1981. Además, publicó obras de enseñanza y difusión del Mapudungun: *Textos orales en mapuche o araucano del centro-sur de Chile* en 1984 y *El mapuche o araucano. Fonología, gramática y antología de cuentos* en 1992.

Otras investigaciones destacadas fueron hechas por la profesora María Katrileo quien participó en la proposición del *Alfabeto mapuche unificado,* 1986 - 1988. Entre sus principales publicaciones se encuentra la gramática normativa *Mapudunguyu. Curso básico de lengua mapuche* de 1987, y el *Diccionario Lingüístico etnográfico de la lengua mapuche* de 1995. La obra de Katrileo representa una de las contribuciones más importantes al esfuerzo por codificar el Mapudungun y dotarlo de una norma estándar. El aporte extranjero ha sido de investigadores como Bryan Harmelink, quien con un enfoque lingüístico funcionalista publicó estudios como *The uses and functions of mew in Mapudungun,* 1987, *The expression of temporal distinctions in Mapudungun,* 1988 y *El hablante como punto de referencia en el espacio: Verbos de movimiento y sufijos direccionales en Mapudungun,* 1990. También publicó textos de divulgación: *Vocabulario y frases útiles en Mapudungun,* 1990 y *Manual de aprendizaje de la lengua mapuche. Aspectos morfológicos y sintácticos,* 1996.

Los aportes a la investigación del Mapudungun en el período contemporáneo vienen también del extranjero y del ámbito nacional, y la especificidad crece, ya que los fenómenos lingüísticos se estudian desde una perspectiva comparativista, lo que impone el ocuparse de puntos específicos de las lenguas, sin embargo no se deja de lado la descripción de la gramática del Mapudungun. Una obra que describe el Mapudungun y a la vez entrega una visión comparativista es *Mapudungun. El habla mapuche* (Zúñiga, 2006), encontramos numerosos ejemplos de otras lenguas del mundo para explicar los fenómenos que se dan en la lengua mapuche, el autor de esta obra, Fernando Zúñiga, publica a menudo estudios sobre diferentes aspectos del Mapudungun y otras lenguas.

Incluiremos en este período el aporte hecho por la Investigadora Ineke Smeets, aunque su obra data de 1989, y se podía consultar, ésta fue oficialmente publicada, previa revisión y actualización, recién el año 2008. También es una gramática descriptiva del Mapudungun, muy valiosa para nuestro proyecto, pues en ella encontramos el sistema de clasificación morfológica ideal para utilizar como base, nos referimos a la ubicación de morfemas por slots, y muy importante también, a la descripción de la combinatoria de los morfemas. Hay otras obras descriptivas de gran valor como *Morfología y Aspectos del Mapudungun* (Lonkon, 2011) y *La lengua mapuche en el siglo XXI* (Katrileo, 2010), que utilizaremos como información de contraste para los aspectos teóricos tratados por Smeets.

Algunos temas estudiados en esta época, con respecto al Mapudungun, son la incorporación nominal (Baker & Fasola, 2006; Baker, 2006; Baker, Aranovich & Golluscio, 2011); las propiedades de los afijos -nge y -le del Mapudungun (Lonkon, 2007), que corresponden en cierta medida a los verbos ser y estar del castellano; la inversión y la marca de objeto diferencial (Zúñiga & Herdeg, 2007); las construcciones seriales verbales (Fernández-Garay & Malvestitti, 2009); operaciones de cambio de valencia (Zúñiga, 2009), etc.

1.2.2. Herramientas computacionales para el Mapudungun

En el ámbito computacional no es muy numeroso el trabajo realizado, lo más destacado es el proyecto *Avenue* del Language Technologies Institute de la Universidad Carnegie Mellon, de Pensilvania, EE. UU. En un documento titulado *Informe Final* (Language Technologies Institute, 2005), se resume el proyecto con sus logros y tareas pendientes. El proyecto se inició en 1999 con el nombre *NICE* (Native-Language Interpretation and Communication Environment), posteriormente *Avenue*. Sus objetivos eran la creación de un sistema de traducción simultánea y contribuir a la preservación y desarrollo de lenguas indoamericanas[14]. En lo referente al Mapudungun el proyecto se realizó entre los años 2000 y 2004 en colaboración con el Instituto de Estudios Indígenas de la Universidad de la Fron-

14 El proyecto Avenue trabajó con tres lenguas estableciendo equipos de trabajo en tres países, Aymara en Bolivia, Quechua en Perú y Mapudungun en Chile.

tera, en Chile. Con los objetivos específicos de crear un grafemario para *Avenue*, generar las reglas gramaticales de acuerdo al grafemario, generar un corpus paralelo Mapudungun/Castellano, recopilar un glosario, programar un corrector ortográfico de Mapudungun para OpenOffice y, construir un prototipo de traductor computacional Mapudungun/Castellano.

El informe dice que todas estas tareas se llevaron a cabo en mayor o menor medida. Interesante es el apartado que habla del traductor automático, un *sistema de traducción basado en ejemplos*[15], sin embargo, los investigadores dicen estar trabajando al momento del informe, en un sistema de traducción basado en reglas de transferencia[16], en el que se han consignado las construcciones básicas del Mapudungun: oraciones simples con verbos intransitivos y transitivos, frases nominales con determinantes y modificadores, frases verbales con distintas especificaciones temporales y aspectuales, frases verbales pasivas, así como frases verbales con concordancia inversa[17]. También es relevante el apartado que trata sobre el analizador morfológico, según la información que se desprende del informe, se trataría de un sistema de FST alimentado por 105 sufijos y unos 700 verbos. El documento no especifica si estos "verbos" son sólo raíces verbales o también hay de otro tipo, como nominales, tampoco cómo funciona en detalle el analizador, por algunos datos que hemos podido recopilar, el analizador esta-

15 Example-Based Machine Translation, EBMT, este método trabaja buscando en su base de datos traducciones completas, o más a menudo porciones de cláusulas, las que han sido traducidas previamente por traductores humanos.

16 En el sistema Transfer-Based machine translation, TBMT, el texto original se analiza primero morfológica y sintácticamente, obteniendo como resultado una representación sintáctica superficial. Esta representación se transforma a continuación en otra más abstracta que hace especial énfasis en aspectos relevantes para el proceso de traducción e ignora otro tipo de información. El proceso de transferencia convierte esta última representación, ligada aún al idioma original, a una representación al mismo nivel de abstracción pero ligada al lenguaje objetivo. Estas dos representaciones son las llamadas normalizadas o intermedias. A partir de aquí el proceso se invierte: los componentes sintácticos generan una representación del texto y finalmente se genera la traducción.

17 La inversión o forma inversa en Mapudungun es una de las dos posibles expresiones del verbo transitivo (la otra es la forma directa) en que hay una tercera persona actuando sobre una primera o segunda, esto se expresa mediante la aparición del sufijo -e antes del modo y -(m)ew al final, siempre que la primera o segunda personas sean pacientes; es decir, no se altera el orden de los constituyentes, que en Mapudungun son sufijos, sino que se agregan y combinan los sufijos detallados para cambiar la direccionalidad de la acción del verbo (Zúñiga, 2006: 114).

ría programado con el lenguaje de programación Perl y también sería capaz de generar oraciones en Mapudungun. Lamentablemente el material del proyecto *Avenue* no se encuentra disponible.

A fines del año 2006 la compañía Microsoft sacó al mercado su producto Windows XP en Mapudungun. Esta no es una herramienta lingüística, pero debido al impacto negativo que tuvo en la comunidad mapuche hemos decidido dar cuenta de ella. Es de suyo conocido el problema que enfrentan las comunidades indígenas con los respectivos gobiernos de los estados en que se encuentran, el pueblo mapuche no es la excepción. Hay un profundo problema político que se arrastra desde mediados del siglo XIX y cuya solución no se vislumbra en el horizonte. Cualquier acción que el gobierno de Chile emprenda, y que tenga que ver con los indígenas, es vista por ellos con desconfianza y recelo, por decir lo menos, y con justa razón, debemos aclarar. Los gobiernos pos-dictadura han facilitado la mercantilización y apropiación a manos de compañías extranjeras de muchos bienes del país, entre ellos los de los indígenas. El gobierno de Chile mediante la Corporación de Desarrollo Indígena (CONADI), intenta legitimar sus acciones en el campo de las relaciones con los pueblos aborígenes, pero ésta es sólo un instrumento del Estado y no un mecanismo real y eficaz para favorecer a los pueblos indígenas. Entre los proyectos patrocinados por la CONADI está la creación de un grafemario mapuche, el Azümchefe (CONADI, 2005), que pretende ser el oficial, moción que la comunidad mapuche rechaza por todo lo expuesto anteriormente; y es justamente este grafemario el que una empresa extranjera como Microsoft utilizó para publicar su versión de Windows XP en Mapudungun, por ello, y las otras razones políticas expuestas, las réplicas de las autoridades espirituales y sociales mapuche no se hicieron esperar (Rolleri, 2006). Elisa Lonkon se expresó al respecto diciendo que el gobierno, a través de la CONADI, con el apoyo de la Universidad de la Frontera, toman una decisión sobre esta lengua "sin ni siquiera consultar a los hablantes". Agrega que lo que aparenta ser una ayuda para los mapuche en el fondo da cuenta de un problema político serio que no se puede dejar pasar: “Se está vendiendo un patrimonio cultural y principal elemento constitutivo del pueblo mapuche a una multinacional". Christian

Martínez Neira, doctor en Sociología e investigador del FONDECYT (Fondo Nacional de Desarrollo Científico y Tecnológico), explica que "el punto no es si en los programas computacionales se puede o no utilizar el Mapudungun. En eso no hay problemas. El asunto es que bajo este pretexto se pretenda oficializar una forma de escritura de este idioma y, de paso, se privatice su léxico". La oposición fue amplia desde la comunidad mapuche, tanto en Chile como en Argentina, sin embargo Microsoft no sólo lanzó su producto, sino que anunció que Microsoft Office también tendría su versión en Mapudungun, que nunca vio la luz. Hoy sólo existe una actualización obsoleta para XP en Mapudungun. De todas maneras, debemos rescatar que detrás del proyecto XP hubo un gran trabajo lingüístico.

Otra herramienta que hemos encontrado es un prototipo de traductor *Winka Süngun a Che Süngun* (Rumian, 2011), una variante del Mapudungun. Este programa traduce conjugaciones verbales del castellano. Las expresiones válidas para traducir comienzan con pronombres personales, pueden contener la negación, los indicadores de transición (me, te, etc.) y verbos conjugados en los cuatro tiempos básicos del castellano (presente, pretérito indefinido, pretérito imperfecto y futuro). El programa reconoce 144 verbos (amar, andar, cantar, jugar, vivir, barrer, etc.), 163 excepciones a las reglas generales de conjugación del castellano y 41 terminaciones verbales. Ejemplos de expresiones válidas: "yo te vi", "te trajo", "yo no viviré","ellas me escuchaban". En palabras de su autor: “Evidentemente éste es un programa experimental que debe ser pulido”.

Por nuestra parte hemos desarrollado el *Norwirin Mapudungun Trapümfe*[18] (Chandía, 2008), una herramienta de unificación ortográfica para el Mapudungun, para poder disponer de textos electrónicos con una ortografía uniforme o con la menor variación posible. El Mapudungun comenzó a ser escrito con el alfabeto latino por los jesuitas de la época colonial para evangelizar a los mapuche. Aún no hay una grafía fijada o normativizada, y desde la primera gramática del siglo XVII hasta nuestros días se han propuesto distintas soluciones ortográficas para esta lengua, unas más utilizadas, otras menos, pero presentes en los diversos documentos escritos en Mapudungun, a veces se encuentran mezcladas, e

18 Disponible en http://www.chandia.net/küdawkawe.

incluso con una ortografía castellanizada.

También debemos mencionar nuestro trabajo previo, donde nació este proyecto. El *Dungupeyem* (Chandía, 2012) fue concebido hace ya un año y medio aproximadamente, y en estos momentos está en estado de prototipo; tiene desarrolladas expresiones regulares que formalizan las formas verbales predicativas transitivas e intransitivas de la lengua mapuche, con un tratamiento inicial de la movilidad de los afijos y la formación de algunos tipos de raíces, aunque funciona correctamente, aún falta mucho para darlo como un producto acabado.

Por último, se puede encontrar por Internet una gran variedad de glosarios y diccionarios, la mayoría Castellano-Mapudungun, Mapudungun-Castellano, pero también algunos Mapudungun-Inglés. En estos momentos estamos formando un grupo para digitalizar y publicar por Internet con acceso gratuito, los diccionarios de Augusta, *Araucano-Español* y *Español-Araucano,* de 1916, los más completos hasta la fecha.

1.2.3. MAGs para lenguas aglutinantes

Hemos acotado el campo de las herramientas lingüístico-computacionales porque son cada vez mayores los esfuerzos puestos en las lenguas minoritarias, por ello hay mucha información al respecto que no podríamos resumir en este documento. Al hacer la siguiente búsqueda en Google "+transducer +morphology +agglutinative", hemos encontrado referencias a las siguientes lenguas: Abkhaz (Turquía), Aymara (Bolivia), Estonio, Faroese (Dinamarca), Finés, Georgiano, Gikuyu (Kenya), Hindi, Húngaro, Indonesio, Inuktitut (Canadá), Japonés, Kannada (India), Kinyarwanda (Rwanda), Coreano, Malagasy (Madagascar), Malayalam (India), Marathi (India), Mongol, Nguni (Sudáfrica), Persa, Quechua (Perú), Saami (Laponia), Sánscrito, Setswana (Sudáfrica), Sotho (Sudáfrica), Tamil (India), Tulu (India), Turco, Turkmen (Turkmenistán), Vasco, Wolof (Senegal), Zulú (Sudáfrica). El Turco y el Finés son las lenguas con más presencia, luego están el Coreano y el Vasco. Las lenguas más próximas al Mapudungun sólo aparecen una

vez cada una, el Aymara y el Quechua[19]. Seguramente hay más trabajo computacional realizado para estas lenguas, pero hemos hecho esta búsqueda para estimar su presencia en el desarrollo computacional actual. Recordemos que el proyecto *Avenue* incluía además de Mapudungun, Aymara y Quechua.

En la última década se ha experimentado con FST, y con sistemas híbridos que los combinan con los métodos estadísticos. Pero es casi unánime la percepción de que las lenguas aglutinantes necesitan de un sistema robusto de análisis morfológico para cualquiera de las tareas posteriores que se quiera realizar, como el análisis sintáctico o la traducción automática; y son los sistemas basados en reglas los más adecuados para efectuar este proceso, sin descartar la integración con otros sistemas. Hay diferentes propuestas para las técnicas a seguir, por ejemplo separar el tratamiento de restricciones morfotácticas secuenciales y no-secuenciales. Las restricciones secuenciales se aplican en la fase de segmentación, y las no-secuenciales en la fase final de combinación de formas. La temprana aplicación de restricciones morfotácticas secuenciales durante el proceso de segmentación hace factible una implementación eficaz del analizador morfológico. Esta propuesta fue aplicada en un MAG para el Vasco, y como resultado se evita una cantidad excesiva de segmentación sin sentido antes del proceso de unificación, computacionalmente más costoso (Aduriz et al., 2000).

Una preocupación constante entre los desarrolladores de herramientas computacionales es la evaluación, válido también para los MAG. Hay propuestas de generación de bancos de pruebas para estos sistemas, que ayudarían a los diseñadores proporcionándoles datos tanto de formación como de evaluación, y permitirían la comparación entre programas (Maxwell, 2002). De igual modo, es constante el intento de minimizar recursos en las diferentes áreas computacionales, como ya habíamos adelantado, la creación de sistemas híbridos es una propuesta concreta al respecto. Debido a que es poco habitual disponer de texto morfológicamente anotado o raíces debidamente identificadas para las lenguas aglutinantes, el procesamiento no supervisado o mínimamente supervisado es

19 En este caso no nos estamos refiriendo a una cercanía genética, puesto que hay dudas al respecto, sino a una cercanía tipológica y geográfica, de hecho, estas tres lenguas estuvieron en contacto durante el período pre colonial.

muy útil y esencial para una rápida y amplia cobertura de la lengua en estudio (Wicentowski, 2002).

Más arriba mencionamos haber encontrado recursos para el Aymara, es un sistema (XFST) del tipo que nosotros estamos desarrollando. Su autor describe cómo la aproximación computacional ayudó a resolver un problema descriptivo sobre un morfema de esta lengua,"los lingüistas descriptivos y computacionales ganan mucho al trabajar juntos" (Beesley, 2003). Para el Coreano, también basado en FST, hemos encontrado un analizador léxico, en rigor es una analizador morfológico, que utiliza un diccionario de lemas, para identificar las raíces, y el sistema de FST para los afijos, de manera de poder devolver definiciones aproximadas dependiendo de la composición de la palabra buscada (Bae & Choi, 2003). Para el Finés hay muchos recursos disponibles, lo que no quiere decir que el problema del análisis morfológico esté resuelto, se siguen buscando métodos más versátiles, como la segmentación morfémica sin supervisión, un método que demanda menos tiempo y menos conocimientos lingüísticos (Creutz & Lagus, 2007). Sin embargo la construcción de sistemas basados en FST continúa siendo el motor de los MAG. A lo largo del tiempo los métodos FST han alcanzado una amplia utilización y también están disponibles en implementaciones de código abierto como OpenFST[20] o FOMA[21] (Koskenniemi, 2008). Un ejemplo de utilización de los recursos disponibles es la construcción de un treebank[22] paralelo entre el Turco y Sueco, para el cual se utilizaron los estudios y herramientas desarrollados hasta la fecha (Megyesi et al., 2008).

La lengua sudamericana que ha recibido mayor atención hasta ahora es el Quechua, para la cual se ha desarrollado un MAG que nos sirve de guía para nuestro propio sistema, y que demuestra que las técnicas de FST son perfectamente capaces de capturar la complejidad morfológica del Quechua (y otras lenguas aglutinantes), una vez que los procesos lingüísticos que determinan la for-

20 http://www.openfst.org/twiki/bin/view/FST/WebHome

21 http://code.google.com/p/foma/

22 Un treebank es una colección de texto sintácticamente anotada, en donde la anotación casi siempre sigue una teoría sintáctica, principalmente basada en estructuras de constituyentes o de dependencias.

mación de palabras han sido desentrañados. El analizador trabaja correctamente en un 95% de los casos, y fue probado con 8 textos de dominios totalmente diferentes, textos legales, poesía y extractos de la Biblia. Una ventaja adicional es la alta eficiencia de las FSN, que permiten procesar textos muy largos en sólo unos segundos (Ríos, 2010).

Para acabar, otro ejemplo de utilización de FST es un diacritizador para el Persa. Debido a la ambigua ortografía del persa es completamente necesario marcar la acentuación de las palabras para que las partes de la oración y los significados léxicos correctos sean encontrados y devueltos como resultado (Nojoumian, 2011), y esto se llevó a cabo exitosamente con el sistema de Xerox, el XFST.

Como se puede ver en este pequeño resumen, las aplicaciones basadas en FST parecen ser las más adecuadas para el análisis y generación morfológicos de lenguas aglutinantes, que además no cuentan con recursos computacionales previos; estas aplicaciones pueden ser optimizadas con otras herramientas de procesos sin supervisión, que simplifican la tarea del desarrollador y ayudan a alcanzar resultados más exactos.

1.3. Objetivos de la tesis

Como hemos planteado en la presentación del tema de investigación, nuestro objetivo es proveer de herramientas computacionales capaces de procesar texto escrito en Mapudungun, y que estas mismas herramientas generen outputs utilizables por otras herramientas lingüístico-computacionales, o directamente por lingüistas interesados en estudiar diferentes aspectos de esta lengua. De esto se desprende que los objetivos son aplicables a tres áreas, la práctica, la descriptiva (teórico-lingüística) y la computacional.

1.3.1. Objetivos prácticos

En esta área de objetivos enmarcamos la aplicación y resultados de la uti-

lización del MAG, es decir, la obtención de material textual en forma de glosas morfológicas, y la obtención de cadenas textuales generadas a partir de estas glosas (análisis y generación morfológicas). Este material generado por el proceso de análisis tiene distintos usos; sirve de paso previo para un tagger morfológico, una aplicación que entrega no sólo la glosa morfológica, sino que también la unidad de la cual se desprende, podríamos obtener algo como:

peñi$_{[hermano]}$-ye$_{[VERB]}$-w$_{[REF]}$-y$_{[IND]}$-u$_{[1NS]}$-Ø$_{[DL]}$

En este caso hay más aplicaciones implicadas aparte del analizador morfológico que ha recibido como input la palabra *peñiyewyu* : tokenizadores, formateadores de texto y el tagger.

También es un proceso previo a los parsers, analizadores sintácticos, y ambos a su vez son procesos cuyos resultados pueden aplicarse a la traducción automática (MT). Sin dejar de lado que las glosas del analizador morfológico, su resultado inmediato, son de gran valor para cualquier estudio lingüístico, del Mapudungun en este caso, llevado a cabo por morfólogos, lexicógrafos u otros especialistas de las lenguas.

En el sentido contrario del proceso, la generación, encontramos el ámbito de la evaluación, que puede ser aplicado tanto a las herramientas que acabamos de mencionar en el párrafo anterior, como a las dudas o experimentos de los propios lingüistas.

1.3.2. Descripción del Mapudungun

Si bien basaremos el desarrollo de nuestro MAG en la descripción de *A Grammar of Mapuche*, el proceso de investigación nos llevará a corroborar, cuestionar o incluso modificar las afirmaciones hechas por Smeets, de hecho, ya estamos al tanto de la discusión que ha propuesto Zúñiga a propósito del sistema agentivo descrito en nuestro libro guía, que otros autores denominan sistema inverso. Esta diferencia de conceptualización del fenómeno nos puede llevar a cambiar la denominación que Smeets dio a los morfemas implicados.

En el paradigma transitivo de conjugación verbal se incluyen las partículas de negación, habiendo una partícula para cada modo, Smeets afirma que la forma negativa para el imperativo *-ki-* combina con las formas Indicativas, en donde agrega la *-l-* del Condicional. Sin embargo, formalmente combina con las formas Imperativas portmanteau y también con algunas conjugaciones Condicionales, de hecho, donde aparece el sufijo *-ki-* aparece obligatoriamente el sufijo de Modo Condicional *-(ü)l-*; para nosotros esta es una interacción morfológica que semánticamente indica un modo diferente, el Imperativo; incluso se da el caso de que allí donde no hay una forma Imperativa afirmativa, hay una negativa.

Un punto que Smeets no trata en su obra es la clasificación de las raíces, identificar tipos de raíces tanto nominales como adjetivales, esto es importante pues no todas las raíces de estos tipos combinan con todos los sufijos verbalizadores, y la diferencia entre ellas, según hemos visto, es semántica, es decir ciertos afijos se unen a raíces de un determinado campo semántico, en esto hay muy poca investigación, y esperamos poder aclarar estos puntos mediante nuestro trabajo.

En el apartado del estado de la cuestión ya mencionamos unas cuantas obras que describen los diferentes aspectos de la lengua mapuche, son los puntos de comparación que utilizaremos para validar el aspecto teórico de nuestro trabajo, y corregir o ampliar su alcance llegado el caso.

1.3.3. Punto de vista computacional

La primera aplicación, y la que demanda más trabajo, es el MAG, el analizador y generador morfológicos. El siguiente paso será construir un guesser morfológico, una herramienta que pueda reconocer la raíz o raíces de las palabras que no están incluidas en los archivos que recopilan los distintos tipos de raíces, para, de esta manera, ir aumentando las entradas del léxico. Una vez alcanzados los objetivos anteriores daremos inicio a la implementación de un corrector ortográfico. Entre otras utilidades que queremos desarrollar están un desambiguador y un etiquetador de las partes de la oración, y si el tiempo nos alcanza, inten-

taremos desarrollar un analizador sintáctico básico, para preparar el camino a las herramientas lingüístico-computacionales más complejas como el analizador sintáctico propiamente tal y un sistema de traducción automático para el Mapudungun.

Nos hemos referido al proyecto *Avenue* en unas cuantas ocasiones, y también hemos comentado que al parecer los frutos de este proyecto no están disponibles para la comunidad lingüística. Mediante nuestro trabajo queremos suplir esta falta, aunque no pretendemos llegar tan lejos como se proponía el equipo del LTI. Esperamos también, que este trabajo sirva no sólo para el Mapudungun sino que aliente a otros investigadores a implementar sistemas similares para otras lenguas minoritarias. Pero sobre todo creemos que nuestro trabajo será un gran apoyo para investigadores del Mapudungun y les ayudará a establecer o cimentar sus teorías lingüísticas con respecto a esta lengua.

Otro punto importante de nuestro trabajo será la investigación sobre herramientas de análisis y generación morfológicas no supervisados, que pueden complementar nuestro proyecto y mejorarlo, haciéndolo aún más asequible y versátil. En este campo hay bastantes aplicaciones pero todas necesitan de ingentes corpus y corpus paralelos para ser entrenadas, en el caso del Mapudungun estos no están disponibles, al menos no en la cantidad y calidad que se necesitaría para establecer un sistema robusto y fiable, por ello nuestro trabajo es fundamental como soporte de estas aplicaciones sin supervisión. Debido a la alta cantidad de posibles formas léxicas, muchas formas perfectamente válidas no se observarán de ningún modo en los datos de entrenamiento, ni siquiera en cantidades grandes de texto. Estos problemas son especialmente severos para lenguas con una morfología muy prominente, como el finés y el turco. Por ejemplo, en finés, un verbo puede aparecer en miles de formas diferentes (Creutz & Lagus, 2007).

Sin embargo, hay otros aspectos en los que nos podemos beneficiar de las aplicaciones no supervisadas, podemos utilizar los guessers morfológicos para obtener raíces automáticamente, y con un algunos algoritmos estadísticos

podemos hacer que no sólo se fragmenten las cadenas textuales en raíces y afijos, sino que además podríamos llegar a clasificar las raíces en sus diferentes tipos, nominales, verbales, adjetivales, etc. Hay otros procesos que también podrían ser ejecutados en la modalidad no supervisada y que nos ayudarían a construir un sistema más completo.

1.4. Marco teórico

Nuestro trabajo se desarrolla a partir de dos núcleos, una descripción de la lengua mapuche que en su aspecto morfológico ubica los sufijos en slots, y un sistema computacional idóneo para el tratamiento de reglas que expresan los fenómenos morfofonológicos de las lenguas.

En la sección del estado de la cuestión nos hemos referido a los trabajos descriptivos acerca de la lengua mapuche, cualquiera de estos trabajos, o todos en su conjunto, nos servirían como base de la implementación computacional, pero es el trabajo de Ineke Smeets, *A Grammar of Mapuche*, el que nos brinda una composición tal que es mucho más fácil plasmar el funcionamiento morfofonológico del Mapudungun en los FST. En efecto, la clasificación de sufijos y su combinatoria están hechos de una manera que en ocasiones sólo hace falta traducir las descripciones a RE. Smeets ubica los sufijos en distintas posiciones o slots y describe las condiciones en que cada sufijo se hace presente en la cadena verbal.

Una forma verbal mapuche consta de una raíz seguida por uno o más sufijos derivativos opcionales y, al menos, un sufijo flexivo. El grupo de marcadores flexivos que se ubican al final de la forma verbal contiene sufijos que indican persona, número, modo, nominalización flexiva, tiempo, aspecto y negación. El Mapudungun tiene alrededor de cien sufijos verbales, dependiendo de la posición relativa y de la función, los sufijos verbales fueron asignados a uno de los 36 slots propuestos, siendo el slot 1 el más alejado de la raíz (Smeets, 2008:149). Los sufijos se presentan siguiendo reglas de coaparición o exclusión que hacen que en la cadena verbal pocas veces coexistan más de 10 sufijos.

En la parte computacional son los FST nuestro punto de apoyo, y el manual de aplicación será el trabajo de Beesley y Karttunen, *Finite State Morphology.* El éxito de la labor depende del entendimiento que logremos de la descripción del Mapudungun y de la habilidad que alcancemos en la creación de RE para interpretarlas y aplicarlas al sistema de análisis y generación morfológicos.

Por otra parte tenemos en mente la incorporación de sistemas no supervisados, entre estos encontramos aquellos que mediante algoritmos de comparación seleccionan las diferentes cadenas léxicas, que en principio corresponderían a diferentes morfemas; sistemas estadísticos más avanzados también están en nuestro punto de mira, pero aún no hemos investigado lo suficiente como para saber cuáles son las aplicaciones que harían a nuestro sistema entregar mejores resultados. Si bien hemos hecho algunas pruebas con los sistemas más sencillos, no estamos en condiciones aún de decidir qué línea tomar con relación a estos sistemas de tratamiento del lenguaje. Sin embargo, hemos comprobado, mediante toda la documentación que hay disponible, que los sistemas como el nuestro se complementan y robustecen con este tipo de herramientas, por lo tanto es un muy importante el tiempo de investigación que dedicaremos a la búsqueda de sistemas que brinden mayor compatibilidad con nuestro MAG y que puedan, mediante su incorporación, formar parte de una estructura de tratamiento de texto escrito en Mapudungun.

1.5. Hipótesis

Creemos que nuestras hipótesis tienen dos vertientes, por un lado la investigación propiamente tal, cuyos puntos de dasarrollo se dividen a su vez en el teórico descriptivo y el computacional. El primer punto, como ya hemos mencionado en ocasiones anteriores, es la clasificación de las raíces nominales y adjetivales para así poder crear una regla de combinatoria con los diferentes morfemas de verbalización que se encuentran en el slot 36. Computacionalmente la investigación nos llevará al campo de las herramientas no supervisadas y su incor-

poración al sistema basado en FST, esperamos poder crear una estructura compleja e interconectada de aplicaciones para el tratamiento del Mapudungun.

La segunda rama es más concreta, y la podemos resumir en tres puntos: Primero, el Mapudungun es un idioma que puede ser formalizado mediante un lenguaje lógico-computacional. Segundo, dicha formalización es implementable en aplicaciones computacionales que pueden procesar lenguaje natural mediante algoritmos programables. Y tercero, esta aplicación será el resultado de nuestra labor.

1.5.1. Formalizar una lengua

En la actualidad sabemos que cualquier lengua puede ser formalizada, expresada en términos metalingüísticos, mediante más o menos trabajo, y es justamente este trabajo a lo que nos referimos cuando decimos que el Mapudungun es una lengua formalizable, reiteramos que nuestra labor será volcar la descripción hecha por Smeets al sistema computacional que estamos desarrollando. La manera en que esta lingüista describió la lengua mapuche nos ahorra el proceso de clasificación y ubicación de los sufijos; en su libro tenemos casi todo el material a punto para construir las RE que entrarán en juego a la hora de codificar las reglas que interaccionan en la realización morfofonológica del Mapudungun. Por lo tanto, más que una hipótesis, es un hecho que formalizar computacionalmente el Mapudungun será un primer paso fundamental para la consecución de nuestro objetivo.

1.5.2. Implementación del sistema

En el punto anterior postulamos la formalización del Mapudungun, en este apartado daremos destino a esa labor. La ciencia computacional ha desarrollado autómatas capaces de generar procesos mediante la definición de los estados que los componen, estos autómatas son los FST, que entre muchos procesos también se aplican al de formación de palabras (cadenas textuales) mediante el tratamiento de los diferentes aspectos morfofonológicos que presentan las len-

guas y que se pueden formalizar, como vimos anteriormente.

En la morfología de las lenguas existen dos problemas centrales para el tratamiento computacional, la composición morfosintáctica de las unidades léxicas, que la morfotáctica se ocupa de expresar; las palabras se componen de morfemas que están sujetos a reglas de realización. Y las alternancias fonológicas y ortográficas. Con la codificación de estos fenómenos alimentaremos un sistema de cálculo de estados finitos que trabajará procesando cada estado dentro de una red de transducción léxica. Los FST tienen la ventaja de poder gestionar millones de caracteres en pocos segundos ocupando muy pocos recursos. En 1996 el Transductor Léxico del Castellano de Xerox contenía más de 46.000 formas léxicas, que podían generar más de 3.400.000 formas flexionadas, aun así ocupaba sólo 3349 kbytes de memoria, aproximadamente 1 byte por forma flexionada (Beesley & Karttunen, 2003).

1.5.3. Dungupeyem[23]

El resultado de todo este proyecto será el *Dungupeyem*, nuestra herramienta computacional para el tratamiento del Mapudungun, un analizador y generador morfológicos que puede crecer y formar parte de otras herramientas más complejas. De esta manera promoveremos la divulgación y preservación del Mapudungun, con un respaldo fiable en el mundo digital. De manera más general, estamos seguros de que aportaremos al ámbito de las lenguas minoritarias un modelo que puede ser reproducido o adaptado. Pero antes de poner a disposición de la comunidad lingüística, y del público en general, el fruto de este proyecto, tendremos que convertir todo el sistema a un modelo de libre distribución, pues la idea es que sea de acceso gratuito, y tal como se estila en el mundo del software libre, también su mecanismo será modificable de manera libre.

23 Dungu= palabra, habla | pe= indica proximidad temporal y/o física con los hechos | ye= indica rasgo constante | m=marca de verbo nominal instrumental, indica instrumento o locación. La traducción sería algo así como “instrumento que siempre se usa para hacer algo con el lenguaje”, para nosotros: “herramienta del lenguaje”.

2. Metodología de investigación

2.1. Metodología adoptada

Como hemos venido explicando a lo largo de este documento, seguiremos las directrices encontradas en el libro *Finite State Morphology* de Beesley & Karttunen para transcribir las reglas descritas en *A Grammar of Mapuche* de Smeets. Cada evento de la morfología del Mapudungun tiene que ser descrito en forma de RE, sólo la práctica hace que tales expresiones sean cada vez más exactas, por tanto el mecanismo de ensayo y error será el adoptado para comprobar la eficacia de cada una de estas reglas. Cada aspecto morfológico que se formaliza mediante una RE se debe probar para confirmar que funciona de manera adecuada, a la vez, se debe controlar que no se provoquen interferencias con las demás RE. Desde el diccionario de Augusta (Augusta, 1916) se han ido copiando las raíces verbales, nominales, adjetivas, etc., que nos sirven para hacer las pruebas; contamos en estos momentos con un pequeño corpus de unas 400 raíces. En ocasiones, las raíces que tenemos ya almacenadas no son las adecuadas para probar las nuevas RE introducidas al sistema, aprovechamos entonces para recopilar raíces nuevas que nos provean de un buen sustento para garantizar la consistencia y eficacia de la RE recién incluída.

Otra metodología será la comparativa, que adoptaremos al revisar los postulados hechos por la autora, para contrastar las descripciones del Mapudungun de otros autores; el objetivo es asegurar el rigor descriptivo de nuestro propio trabajo. Como hemos comentado, Zúñiga hace algunos reparos tanto a alguna de la terminología empleada por Smeets, como a la explicación teórica de algunos aspectos de la morfología del Mapudungun; además, haciendo una revisión superficial de otras descripciones de la lengua mapuche, hemos constatado

que existen algunas variaciones con respecto a lo que Smeets recopiló, no olvidemos que ella trabajó sólo sobre la variedad dialectal del Mapudungun central (Smeets, 2008:16).

Tendremos que tomar una actitud puramente investigativa al intentar definir las reglas de combinación de las raíces nominales y adjetivales con los diferentes sufijos verbalizadores, en este tema hay muy poca investigación, de hecho Smeets apenas menciona ciertas distinciones semánticas para algunos tipos de raíces verbales, pero en cuanto a las nominales y adjetivales no se encuentra tal descripción. Otros autores tampoco mencionan este aspecto, por tanto tendremos que indagar y sacar conclusiones por cuenta propia, esperando así hacer un aporte al estudio de la gramática del mapudungun.

En la etapa de investigación de las herramientas computacionales no supervisadas, haremos una búsqueda y recopilación del material disponible y, dependiendo de cada caso, haremos las pruebas como indican sus autores. En general los sistemas no supervisados deben ser alimentados con corpus sin anotar, que al ser procesados entregan los resultados para los cuales sus algoritmos están preparados; en el caso específico de los sistemas dedicados al aprendizaje morfológico, el resultado será un listado de raíces y afijos, dependiendo de la complejidad del sistema empleado, los resultados pueden ser más o menos detallados, por ejemplo el sistema podría entregar un listado de raíces, otro de prefijos y otro de sufijos. Algunos sistemas permiten cierto grado de asistencia, con lo que se podría llegar a obtener un listado de raíces nominales, otro de raíces verbales, etc. Seleccionaremos una serie de los algoritmos disponibles dependiendo de las ventajas que presenten con respecto de otros sistemas, y a éstos aplicaremos un corpus especialmente preparado para la fase de pruebas, el sistema que entregue mejores resultados, y tenga mayor compatibilidad con nuestro MAG, será el escogido para implementar dentro del proyecto *Dungupeyem*.

2.2. Herramientas y recursos necesarios para desarrollar el plan de investigación

En la primera etapa del proyecto serán necesarios, a parte de los dos trabajos que hemos mencionado repetidamente, *Finite State Morphology* y *A Grammar of Mapuche*, el software de FST publicado por la compañía Xerox: XFST, un corpus léxico, que durante la implementación de las RE se va construyendo a medida que se prueban las formalizaciones. El poblamiento del corpus, es una tarea que se completará con posterioridad a este paso, como se explica más adelante, en el punto 3.1.3. También será necesario material léxico, de momento tenemos una colección bastante amplia en varios formatos, esto lo tenemos que estandarizar, es decir, procesar para que pueda ser utilizado como input lingüístico a la hora de hacer las pruebas, y que pueda ser etiquetado en el momento en que ya estemos preparado para ello. En términos de cantidad de palabras, contamos con alrededor de unas sesenta mil formas léxicas.

En la etapa de comparación y complementación de la descripción de Smeets, necesitaremos los trabajos de los otros autores que se abocaron a la misma tarea. Nos referimos a los trabajos de Zúñiga, Salas, Lonkon, Katrileo, Oyarzo, Baker y otros que hemos consignado en la bibliografía.

Y, finalmente, para la fase de investigación sobre herramientas no supervisadas, con un acceso a Internet tendremos para comenzar; luego depende de los requisitos que tenga cada una de las aplicaciones que queramos probar. De todas maneras tenemos algunas en mente, el sistema desarrollado por Oliver, pormenorizado en su tesis doctoral (Oliver, 2004) , los explicados por Creutz & Lagus en *Unsupervised Discovery of Morphemes*, del que ya dimos cuenta en el punto 1.2.3. Algunos de los sistemas que probaremos son *Lingüística*[24], *Conexor*[25]

24 http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000

25 http://www.wjh.harvard.edu/soc_help/cnxfdgunix.pdf

y *Automorphology*[26], entre otros, además de los que vayamos descubriendo en nuestro proceso de búsqueda.

3. Plan de trabajo

3.1. Paquetes de trabajo en que se divide la investigación

3.1.1. Completar tratamiento del verbo. Curso de FST. Formalizar frases subordinadas

La primera tarea será completar y afinar el MAG, es decir, terminar de volcar la fenomenología morfológica del verbo, incorporar los casos que tienen que ver con la movilidad de los afijos, las variaciones fonológicas y/u ortográficas. Continuaremos con la formalización de las frases subordinadas, que en vez de utilizar los morfemas de modo, persona y número, utilizan los de flexión nominativa. Durante esta etapa asistiremos por tres meses a un curso sobre FST dictado en la Universidad de Zürich, el curso dura 6 meses en total, pero aún no tenemos financiamiento para la totalidad, estamos solicitando lo que corresponde a los tres primeros meses. La intención es profundizar los conocimientos al respecto y así poder mejorar nuestro trabajo.

3.1.2. Incorporación de frases nominales, adjetivales, adverbiales y otras partes de la oración

En seguida se incorporará la flexión y derivación nominal, trataremos la formación de unidades nominales, siguiendo con las unidades adjetivales y adverbiales. Tendremos que dedicar algún tiempo a la investigación de la cmbinatoria de las raíces nominales y adjetivales con los diferentes sufijos verbalizadores,

26 http://linguistlist.org/issues/9/9-1450.html, http://humanities.uchicago.edu/faculty/goldsmith/Automorphology/

para tratar luego las adposiciones, pronombres, conjunciones y otras partículas de la lengua.

3.1.3. Poblamiento del corpus léxico (raíces)

Continuaremos con el poblamiento de los corpus léxicos, los archivos que contienen los diferentes tipos de raíces, probablemente esta será una tarea semiautomática, en donde intentaremos aprovechar nuestra colaboración con el proyecto de digitalización de los diccionarios publicados por Augusta en 1916.

3.1.4. Investigación e implementación del guesser morfológico

En la fase siguiente nos dedicaremos al guesser morfológico, herramienta fundamental para alimentar los corpus léxicos de manera automática. Tenemos un par de alternativas que ponderaremos llegado este momento, crear un guesser mediante FST, o implementar uno no asistido como el propuesto por Oliver en su *Tesis Doctoral* (Oliver, 2004). En esta etapa también investigaremos otras aplicaciones que utilicen técnicas no asistidas, y que podamos hacer interactuar con nuestro MAG, ya sea para mejorarlo, complementarlo o extenderlo.

3.1.5. Conversión de código y publicación web

Finalizado todo lo anterior nos dedicaremos a la transformación del código del MAG en uno de código abierto, para ponerlo a disposición del público en general vía web.

3.1.6. Desarrollo de otras herramientas: etiquetador morfológico, corrector ortográfico y shallow parser

Las herramientas que desarrollaremos a partir de este momento son un etiquetador morfológico y un corrector ortográfico, finalmente intentaremos avanzar lo más posible en el camino hacía el análisis sintáctico.

3.1.7. Extensión a otros dialectos del Mapudungun

Queremos reservar la última etapa para extender el trabajo hecho a partir de *A Grammar of Mapuche*, en donde se describe sólo una variante dialectal, a otras formas del Mapudungun, y generar una cobertura más amplia e inclusiva.

3.2. Calendarización (cronograma)

<table>
<tr><th>Sep</th><th>Oct</th><th>Nov</th><th>Dic</th><th>Ene</th><th>Feb</th><th>Mar</th><th>Abr</th><th>May</th><th>Jun</th><th>Jul</th><th>Ago</th></tr>
<tr><td colspan="12">2013 – 2014</td></tr>
<tr><td colspan="3">Completar tratamiento del verbo</td><td></td><td colspan="3">Curso FST</td><td colspan="4"></td></tr>
<tr><td colspan="3"></td><td colspan="3">Formalizar frases subordinadas</td><td colspan="6">Incorporación de frases nominales, adjetivales, adverbiales y otras partes de la oración</td></tr>
<tr><td colspan="12">2014 – 2015</td></tr>
<tr><td colspan="3">Poblamiento del corpus léxico (raíces)</td><td colspan="4">Investigación e implementación del guesser morfológico</td><td colspan="3">Conversión de código y publicación web</td><td>Desarrollo de otras</td></tr>
<tr><td colspan="12">2015 – 2016</td></tr>
<tr><td colspan="4">herramientas: etiquetador morfológico, corrector</td><td colspan="8">Investigación y desarrollo de un shallow parser (analizador sintáctico superficial – prototipo)</td></tr>
<tr><td colspan="12">2016 – 2017</td></tr>
<tr><td colspan="5">Extensión a otros dialectos del Mapudungun</td><td colspan="7">Redacción de la tesis</td></tr>
<tr><th>Sep</th><th>Oct</th><th>Nov</th><th>Dic</th><th>Ene</th><th>Feb</th><th>Mar</th><th>Abr</th><th>May</th><th>Jun</th><th>Jul</th><th>Ago</th></tr>
</table>

4. Bibliografía

4.1. Bibliografía comentada

En cualquier estudio que se haga sobre el Mapudungun es inevitable hablar de su morfología. Es el caso del trabajo que hemos identificado como nuestro libro guía, *A Grammar of Mapuche*, y de otros que son fundamentales para la compresión de la estructura del idioma mapuche. También incluimos algunos artículos que revisan el trabajo de Smeets y aportan algunas consideraciones sobre su investigación. No haremos una lista detallada de todos los trabajos al respecto de la morfología del Mapudungun pero mencionaremos algunos de relevancia.

4.1.1. Some notes on the Mapudungun evidential (Zúñiga, 2003)

Comencemos pues con este artículo en donde se hace un análisis de la partícula -rke, un morfema que Smeets etiqueta como REP, reportative. Zúñiga repasa las definiciones dadas por Augusta (1903), Mösbach (1962), Smeets (1989)[27] y Salas (1992)[28], para mostrar que el campo de acción de este morfema es más amplio de lo que describe Smeets. Estudios como este nos permiten afinar nuestro sistema, pues al dilucidar ciertos aspectos que la autora de nuestro libro guía no trató, podemos ampliar nuestro campo de acción.

4.1.2. Gramática Básica de la Lengua Mapuche (Hernández, Ramos y Wenchulaf, 2006)

Mencionamos esta gramática más que por su aporte teórico, por la difusión que se le dio en Chile. Fue publicada por la CONADI (Corporación Nacional

27 Zúñiga toma como referencia el documento de la tesis de Smeets de 1989, nosotros la versión revisada y publicada del año 2008.

28 Nosotros contamos con la edición del año 2006 del mismo libro, a cargo de Fernando Zúñiga.

de Desarrollo Indígena), una organización gubernamental encargada de tratar los temas indígenas. En su parte III se trata el *Plano Morfosintáctico*, encontramos aquí la *Clasificación de los morfemas: Raíz, Prefijos, Sufijos Inflexionales y Sufijos Derivacionales*. En la parte IV encontramos la *Morfosintaxis del Mapuzungun*[29], que incluye *El Sustantivo*, apartado en que se hace mención a la formación de frases nominales a partir de la raíz nominal y la sufijación de ciertos morfemas, para terminar con la composición, la unión de raíces nominales. Lo propio se hace en la parte dedicada al verbo. Este libro es un buen comienzo para quien no tiene conocimiento previo de esta lengua. También para los interesados que no tienen conocimientos lingüísticos adelantados.

4.1.3. Mapudungun. El habla mapuche (Zúñiga, 2006)

El año 2006, Fernando Zúñiga publica una descripción muy completa del Mapudungun, y meses más tarde dirige la reedición del libro de Adalberto Salas: *El mapuche o araucano,* del año 2006, publicado originalmente en 1992. Zúñiga es un comparativista muy activo que actualmente trabaja en Suiza, desde donde constantemente publica artículos sobre diferentes aspectos del Mapudungun. En su obra, el lingüista parte por ubicar al Mapudungun en el concierto mundial de las lenguas; en seguida, en el capítulo I, con una visión sociocultural, revisa aspectos del pueblo mapuche. A partir del capítulo II entra en materia lingüística con los aspectos fonético-fonológicos de la lengua mapuche y, como suele hacerse en la mayoría de los estudios sobre Mapudungun, sino en todos, se menciona el problema de la escritura y las diferentes propuestas para escribir esta lengua.

En el capítulo III encontramos los temas que se relacionan más directamente con nuestro trabajo. Este capítulo, el más extenso, se titula *Las palabras del Mapudungun*, y en él se trata la morfología del sustantivo y del verbo, contiene muchas notas comparativas que ayudan al lector a situar y comprender los fenómenos tratados; los ejemplos de otras lenguas tienen la virtud de ampliar el

29 La CONADI también impulsó la creación de un grafemario mapuche en donde lo que corresponde a la grafía "d" de otros grafemarios, corresponde a la "z" en el Azümchefe, el grafemario de la CONADI. Esta variante se debe a que la pronunciación de la "d" no es apical en el Mapudungun tradicional sino interdental, unas veces con una realización sonora ð, y otras sorda θ, por ello la representación con la grafía "z".

campo visual y por tanto de facilitar la comprensión del objeto de estudio. De manera detallada, Zúñiga nos enseña la estructura morfosintáctica de las frases nominales y verbales, también los aspectos relacionados con pronombres, adjetivos, adverbios, adposiciones y otras partes de la oración. Es muy interesante un apéndice que el autor incluye en este capítulo titulado *El perfil tipológico de las palabras mapuche*, en donde explica y discute la clasificación tipológica hecha respecto del Mapudungun desde dos puntos de vista, el de la tipología morfológica tradicional y el de una visión más renovada sobre la tipología lingüística.

En el capítulo IV, *Las oraciones del Mapudungun*, podemos leer sobre las oraciones sin predicado verbal; la transitividad y el orden sintáctico en oraciones simples; las cláusulas condicionales, causales y atributivas de las oraciones compuestas; las oraciones interrogativas; y un apéndice dedicado al discurso indirecto. En los capítulos finales se ofrece una colección de textos mapuche que incluyen epew[30], nütram[31] y algunos poemas de Leonel Lienlaf[32] en versión bilingüe y comentada. Dos breves glosarios, uno Castellano – Mapuche, y otro Mapudungun – Wingkadungun[33]. El libro culmina con una bibliografía extensa y actualizada, una breve lista de términos lingüísticos y el índice de los archivos de audio, porque este libro además cuenta con un disco compacto que contiene 85 pistas

30 El epew es una narración, generalmente de argumento ficticio, que tiene la función de entretener y enseñar, o inculcar valores positivos como la honradez, la lealtad, la justicia, etc. Los protagonistas son a menudo animales que representan los rasgos del carácter humano que se busca tematizar. (Zúñiga, 2006: 267-268).

31 En el nütram se narran sucesos que se consideran verídicos y tienen como protagonistas típicos a los antepasados, en general a miembros fallecidos de la comunidad y en ocasiones a personas vivas. Los eventos narrados van desde las experiencias personales hasta acontecimientos remotos en el tiempo, algunos de los cuales pueden considerarse legendarios desde la perspectiva occidental. (Zúñiga, 2006: 267).

32 Leonel Lienlaf nació en la comunidad de Alepwe en 1969, el día 23 de junio, coincidiendo con el año nuevo mapuche. Forma parte de una nueva generación de poetas bilingüe que escriben en Mapudungun y Castellano. Ha recibido varios premios y reconocimientos por sus trabajos; entre ellos el Premio Municipal de Literatura de Santiago en 1990. Sus poemas son parte de varias antologías y han sido traducidos al Inglés. De su obra impresa, Se ha despertado el ave de mi corazón (1989) es su primer libro y también es coeditor de Voces Mapuches (2002) publicado por el Museo Chileno de Arte Precolombino. En 1998 realizó un disco compacto, Canto y poesía mapuche, financiado por Embajada de Finlandia.

33 La palabra mapuche *wingka* significa extranjero, invasor; como los extranjeros que llegaron en la época de la conquista eran españoles, generalmente se aplica el término *wingka* a cualquiera de habla castellana, incluidos los chilenos, de ahí que al castellano se le llame *winkadungun* en Mapudungun, es decir “el habla de los invasores”.

con numerosos ejemplos de pronunciación, estructuras oracionales de diversa complejidad, un relato tradicional o epew y algunos poemas de Leonel Lienlaf. Las voces son de Lienlaf y Clara Antünao.

4.1.4. *El Mapuche o Araucano. Fonología, Gramática y Antología de Cuentos (Salas, 1992; 2006)*

El libro de Salas, que ya presentamos al hablar del trabajo de Zúñiga, lo definiremos tomando prestadas las palabras del editor: "Este libro, como puede apreciar el lector echando un vistazo al índice, tiene relativamente menos lingüística "dura" y no incluye un glosario, pero contiene bastante más información histórico-etnográfica. Tiene, asimismo, textos bilingües sustancialmente más numerosos y variados, los cuales están precedidos y/o seguidos de valiosos comentarios del autor". No entraremos en mayores detalles debido a que, como el anterior, también trata los mismos temas, pero listaremos el índice para hacer una idea: Cap. I: *El marco histórico-etnográfico. Los mapuche y su lengua*. Cap. II: *Lingüística Mapuche (Caracterización tipológica)*. Cap. III: *Fonología*. Cap. IV: *El Nombre*. Cap. V: *El Verbo Finito*. Cap. VI: *Los Sufijos de Persona*. Cap. VII: S*ufijos Adverbiales*. Cap. VIII: *El Verbo No Finito*. Cap. IX: *Los Temas Verbales*. Cap. X: *El Género Narrativo*. Cap. XI: *Cuentos Mitológicos*. Cap XII: *Cuentos de Difuntos*. Cap. XIII: *Cuentos de Brujos*. Cap. XIV: *Cuentos de Animales*. Cap XV: *Cuentos Hispánicos*. Epílogo: *El Mapuche: ¿Lengua o Dialecto?*

4.1.5. Baker, 2006

Mark Baker, de la universidad Rutgers de Nueva Jersey, es un lingüista que en la década del 2000 ha hecho mucha investigación con respecto a la composición en varias lenguas, prestando gran atención al caso del Mapudungun. Citamos aquí dos de sus publicaciones del año 2006, la primera, que publica junto a Carlos Fasola lleva un título muy general: *Araucanian: Mapudungun*, sin embargo es un artículo en donde describe ampliamente la composición en lengua mapuche. Este tema es de especial interés para nuestro trabajo, pues pretendemos configurar una herramienta lo suficientemente robusta como para reconocer,

analizar, las raíces verbales o nominales del Mapudungun de manera correcta, y al ser la composición un proceso tan prolífico en la lengua mapuche, es imperativo conocer bien el proceso de la composición de este idioma. En este artículo se discuten las composiciones Verbo+Nombre, Nombre+Nombre, Verbo+Verbo y otras variantes menos productivas.

En el segundo artículo, *Is Head Movement Still Needed for Noun Incorporation? The Case of Mapudungun*, Baker compara uno de sus postulados sobre el movimiento sintáctico del núcleo en la incorporación del sustantivo, con las teorías no lexicalistas de la incorporación del sustantivo. Recalca que el enfoque del movimiento sintáctico captura hechos importantes sobre este fenómeno en el Mapudungun, mientras que las otras teorías los dejan sin explicación, por tanto, dice Baker, este enfoque aún es necesario en el la teoría generativista.

4.1.6. A Grammar of Mapuche (Smeets, 2008)

A pesar de que nos hemos referido innumerables veces al trabajo de Smeets, haremos un pequeño resumen de su contenido, después de todo es el libro que sustenta nuestro proyecto. Esta es la primera gramática de referencia del Mapudungun escrita de acuerdo con los estándares de la lingüística moderna, lo que es un gran avance respecto de las descripciones previas. El libro se divide en nueve partes. La primera parte es una breve introducción sobre el pueblo mapuche, su lengua, el Mapudungun, y del libro mismo. Desde la segunda a la séptima partes encontramos la descripción gramatical del Mapudungun Central, a estos capítulos se añade un apéndice con los paradigmas verbales. La fonología y morfología de esta lengua aparecen en la segunda parte. En la tercera, encontramos la morfología nominal y una primera sección sobre la estructura de las frases nominales. Una segunda sección sobre las frases nominales y las frases sin predicado verbal vienen contenidas en la cuarta parte. Pasando a la quinta parte, nos encontramos con el área más compleja de la gramática mapuche, la morfosintaxis del verbo. En la sexta parte se listan y describen las partículas del Mapudungun, mientras que la séptima parte profundiza en la sintaxis, básicamente en el orden de los constituyentes y en la relación clausal.

En la octava parte hay algunos cuentos y canciones mapuche; en la novena está el breve diccionario Mapudungun-Inglés. Smeets basa su descripción gramatical en su propia investigación y entrevistas con hablantes nativos entre 1977 y 1981, en Chile y Holanda. En la gramática de Smeets encontramos de una manera concisa y clara el tratamiento de las partículas del discurso, y de las formas finitas y no finitas del verbo. La cantidad de datos que contiene este trabajo, lo hacen un recurso muy valioso para los especialistas, los americanistas, los tipologistas (Zúñiga, 2009b) y los lingüistas computacionales.

4.1.7. Finite State Morphology (Beesley & Karttunen, 2003)

Introduciéndonos en la ámbito computacional, tenemos que hablar del trabajo de Beesley & Karttunen, porque es el otro soporte de nuestro proyecto. El libro es una guía de referencia para los sistemas de FST desarrollados por la compañía Xerox. Es una introducción a las técnicas generales de sistemas para el análisis morfológico, que comienza con la historia del propio libro y breves indicaciones para utilizarlo. Luego vienen 9 capítulos con el temario principal, dos apéndices con ejercicios y sus soluciones, y una lista de referencia. También se incluye un CD-ROM con el programa, que se mantiene actualizado en la web del libro: http://www.fsmbook.com

En el primer capítulo, una introducción ligera, se exponen los conceptos primordiales con analogías y ejemplos; siempre evitando un lenguaje demasiado técnico, para hacer más intuitiva la interiorización de los conceptos. De la misma manera se explican las características y aplicaciones de las FSN. El siguiente capítulo, una introducción más técnica, contiene definiciones más formales de lenguaje, relaciones, FSN, RE y otros conceptos. Lenguajes y relaciones pueden formalizarse a través de RE y codificarse mediante FSN, autómatas y transductores, respectivamente. Se introducen los operadores de las RE y se discuten las propiedades de las FSN.

The xfst Interface, es el tercer capítulo, un extenso manual del interprete de comandos o scripts, diseñado para crear y manipular FSN. Enseña la sintaxis

necesaria para implementar las RE y definir los alfabetos multicaracteres para compilarlos y almacenarlos como variables o archivos. También informa de un mecanismo que permite apilar transductores y así ejecutarlos en cascada. En el cuarto capítulo se introduce el lenguaje lexc, un formalismo alternativo para definir FSN, que es puramente declarativo. Avanzando al siguiente capítulo encontramos importantes consejos para una organización efectiva del trabajo que implica construir modelos lingüísticos de estados finitos. Se mencionan aspectos ingenieriles como el control de versiones y la modularidad, y aspectos lingüísticos como la implementación de diferentes ortografías y restricciones léxicas.

El sexto capítulo trata sobre el funcionamiento y evaluación de las herramientas de NLP, de cómo encontrar malfuncionamientos y errores en el sistema. Por otra parte se introducen las utilidades de tokenización y búsqueda, que se ejecutan cada vez que el sistema entra en funcionamiento. Una extensión del cálculo de estados finitos se presenta en el capítulo siete, las etiquetas diacríticas (Flags Diacritics), que permiten relaciones a distancia dentro del paradigma de los estados finitos, para cumplir las dependencias de los fenómenos morfofonológicos sin sobrecargar la FSN, ni hacerla más compleja. La morfotáctica no concatenativa se presenta en el capítulo ocho y propone un sistema léxico eficiente para tratar fenómenos problemáticos de la morfología como la reduplicación o la interdigitación. Para esto se recurre a la noción de que ambos lenguajes del sistema de dos niveles, el superficial y el abstracto (Koskenniemi, 1983), pueden ser interpretados como RE, y recompilados en una FSN más compleja que utiliza el comando de compilación y reemplazo que formaliza una expresión regular como un nuevo input del lenguaje tratado. En el capítulo final se tratan con mucho más detalle las utilidades de tokenización y búsqueda, y se explica cómo construir herramientas más sofisticadas como tokenizadores avanzados, guessers morfológicos, correctores ortográficos, parsers superficiales y otras, mediante la adjunción de FSN y el diseño de diferentes estrategias de búsqueda (Smrž, 2004).

4.1.8. *How to Build an Open Source Morphological Parser Now (Koskenniemi, 2008)*

En la publicación de la Universidad de Uppsala, *Studia Linguistica Upsaliensia 7*, del 2008, Koskenniemi publica este artículo que parte con una breve reseña histórica de los sistemas y desarrolladores de aplicaciones de NLP. Explica cómo se procedía en los años 70 para programar un FST, y que estos eran bidireccionales en la teoría, pero en la práctica solían fallar o no funcionar de esta manera, los procesos de análisis y generación no siempre eran el mismo sistema en direcciones inversas. Más adelante hace un recuento de las herramientas y facilidades de que disponemos hoy, menciona que además de la increíble velocidad y capacidad que tienen los ordenadores modernos, se cuenta con inmensas cantidades de datos y gran cantidad de entidades que se dedican a almacenarlos y clasificarlos, muchas de ellas con fines lingüísticos, además, casi todo está disponible a través de Internet. En cuanto a los mecanismos de FST, recalca que se han perfeccionado, sofisticado y distribuido a lo largo de los centros lingüístico-computacionales del mundo, ganando mucho prestigio, y actualmente también los tenemos disponibles en código abierto. No deja de lado, aunque no profundiza, los sistemas no supervisados, sólo menciona el gran aporte que están trayendo al campo de la lingüística computacional. En lo siguiente toca aspectos de los derechos de copia y licencias de uso, para enlazarlo con los aspectos del código abierto.

En seguida pasa a comentar aspectos de su propio trabajo y proyectos, en especial del *Open Source Morphology for Finnish*, cuyo objetivo es producir un parser morfológico que pueda ser modificado libremente con propósitos de investigación y que a la vez pueda ser utilizado como parte de aplicaciones de código abierto. Le siguen las herramientas que se podrían desarrollar a partir de este trabajo y su aplicabilidad a una infraestructura que prepara la Unión Europea llamada CLARIN, *Common Language Resource and Technology Infrastructure*.

4.1.9. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production (Koskenniemi, 1983)

Debido a la importancia que tiene el trabajo de Koskenniemi para el desarrollo de sistemas morfológicos basados en FST, y porque en sus páginas encontramos los conceptos básicos para entender el funcionamiento de autómatas, transductores y redes de estados finitos, hacemos aquí una reseña de su trabajo de 1983. Aquí se presenta el modelo lingüístico implementado para análisis y síntesis que incorpora un formalismo general para hacer descripciones morfológicas de lenguajes naturales, y una aplicación que no depende de la lengua en estudio, para implementar este modelo. El formalismo de dos niveles está contenido en la aplicación, y está basado en un sistema léxico y un conjunto de reglas de dos niveles. Las reglas están en paralelo y posibilitan un sistema totalmente bidireccional, tanto conceptualmente como computacionalmente. Puede ser interpretado como un modelo morfológico de reconocimiento y generación de palabras.

4.2. Fuentes bibliográficas

Antony P J & Dr. Soman K P (2012). Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey.
En International Journal of Scientific & Engineering Research Volume 3, Issue 3, March-2012.

Aduriz, I. et al. (2000). A word-grammar based morlphological analyzer for agglutinative languages.
Dept. de Lenguajes computacionales y Sistemas, Universidad del País Vasco & Universidad de Barcelona.

Avramidis, E. & Kuhn, J. (2009). Exploiting Xle's Finite State Interface in Lfg-Based Statistical Machine Translation.
Departamento de Lingüística, Universidad de Postdam, Alemania. CSLI Publications. Disponible en: http://csli-publications.stanford.edu

Bae, S. M. & Choi, K. S. (2003). Lexical Analysis of Agglutinative Languages Using a Dictionary of Lemmas and Lexical Transducers.
División de Ciencia Computacional, Dept. de EECS, Instituto Coreano Avanzado de Ciencia y Tecnología.

Baker, M. (2006). Is Head Movement Still Needed for Noun Incorporation? The Case of Mapudungun.
Universidad Rutgers. Nueva Jersey, EE. UU.

Baker, M. & Fasola, C. (2006). Araucanian: Mapudungun.
Universidad Rutgers. Nueva Jersey, EE. UU.

Baker, M.; Aranovich, R. & Golluscio, L. (2011). Two Types of Syntactic Noun Incorporation: Noun Incorporation in Mapudungun and its Typological Implications.
Proyecto Muse: http://muse.jhu.edu

Beesley, K. (2003). Finite-State Morphological Analisys and Generation for Aymara.
Centro de Investigación Xerox Europa. Meylan, Francia.

Beesley, K. & Karttunen, L. (2003). Finite State Morphology.
Publicaciones CSLI. EE. UU.

Chandía, A. (2008a). NMT - Norwirin Mapudungun Trapümfe.
Disponible en: http://www.chandia.net/küdawkawe

Chandía, A. (2008b). Raíces del Mapudungun.
Trabajo de asignatura Lingüística de Corpus, Máster Ciencia Cognitiva y Lenguaje. Universidad de Barcelona, Cataluña.

Chandía, A. (2012). Dungupeyem_alfa_v0.1: un prototipo de analizador morfológico para el Mapudungun a través de transductores de estados finitos.
Tesis de Máster Ciencia Cognitiva y Lenguaje. Universidad de Barcelona, Cataluña.

CONADI (2005). Azümchefe . Hacia la Escritura del Mapuzugun.
Gobierno de Chile, Corporación de Desarrollo Indígena, CONADI. Unidad de cultura y educación.-Subdirección Nacional Sur , Temuko.

Creutz, M. & Lagus, K. (2007). Unsupervised Discovery of Morphemes.
Centro de Investigación de Redes Neuronales. Universidad Tecnológica de Helsinki, Finlandia.

De Augusta, F. J. (1916). Diccionario Araucano – Español y Español – Araucano.
Imprenta Universitaria. Santiago, Chile. Disponible en: http://archive.org/stream/diccionarioarau-c01fluoft

Fernández-Garay, A. & Malvestitti, M. (2009). Las construcciones verbales seriales en mapuche.
Lexis, revista de lingüística y literatura, Vol. XXXII. Departamento de Humanidades. Fondo Editorial. Universidad Católica del Perú.

Gasser, M. (2009). Semitic Morphological Analysis and Generation Using Finite State Transducers with Feature Structures.
Universidad de Indiana, Escuela de Informática y Computación. EE. UU.

Gasser, M. (2011a). Quechua - Spanish AntiMorpho 1.2 User's Guide.
Universidad de Indiana, Escuela de Informática y Computación. EE. UU.

Gasser, M. (2011b). HornMorpho 2.2: User's Guide.
Universidad de Indiana, Escuela de Informática y Computación. EE. UU.

Gasser, M. (2011c). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya.
Conferencia sobre Tecnología para el desarrollo del lenguaje humano, Alejandría, Egipto. Disponible en: ftp://ftp.cs.indiana.edu/pub/gasser/hltd11.pdf

Gasser, M. (2011d). L3 Morpho1.0 User's Guide.
Universidad de Indiana, Escuela de Informática y Computación. EE. UU.

Gasser, M. (2011e). Computational Morphology and the Teaching of Indigenous Languages.
En Proceedings of the First Symposium on Teaching Indigenous Languages of Latin America. CLACS & MLCP, Indiana University Bloomington & Association for Teaching and Learning Indigenous Languages of Latin America (ATLILLA). Indiana, EE. UU.

Hernández, A.; Ramos, N. & Wenchulaf, R. (2006). Gramática básica de la lengua mapuche.
Tomo I.
Temuko: Editorial UC Temuko.

Jurafsky, D. & Martin, J. (2006). Draft: Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.
Publicado por Alan Apt. Nueva Jersey, EE. UU.

Kañumil, T. (2007). Mapucezugun Ñi Cumgeel. Descripción de la lengua Mapuche.
Agrupación Mapuche Witraleyiñ, Florencio Varela, Buenos Aires, Pwelmapu.

Karttunen, L. (2000). Applications of Finite-State Transducers in Natural Language Processing.-
Centro de Investigación Xerox Europa. Meylan, Francia.

Katrileo, M. (2010). La lengua mapuche en el siglo XXI.
Facultad de Filosofía y Humanidades, Universidad Austral de Chile. Valdivia, Chile.

Koskenniemi, K. (1983). Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.Departamento de Lingüística General. Universidad de Helsinki. En Publications No. 11. Finlandia.

Koskenniemi, K. (2008). How to Build an Open Source Morphological Parser Now.
Departamento de Lingüística General. Universidad de Helsinki. En Resourceful Language Technology: Acta Universitatis Upsaliensis. Studia Linguistica Upsaliensia 7. Universidad de Uppsala, Suecia.

Language Technologies Insitute (2005). Proyecto Avenue/Mapudungún: Desarrollo de herramientas informáticas para el mapudungún que se habla en Chile. Informe Final.
Universidad Carnegie Mellon. Pensilvania, EE. UU.

Lonkon, E. (2007). Las propiedades de los afijos -nge y -le del Mapudungun.
Facultad de Humanidades, Universidad de Santiago de Chile. An. Antrop., 41-II.

Lonkon, E. (2011). Morfología y aspectos del Mapudungun.
Universidad Autónoma Metropolitana. México D. F., México.

Megyesi, B. et al. (2008). Supporting Research Environment for Less Explored Languages: A Case Study of Swedish and Turkish.
Departamento de Lingüística y Filología. Universidad de Uppsala. En Resourceful Language Technology: Acta Universitatis Upsaliensis. Studia Linguistica Upsaliensia 7. Universidad de Uppsala, Suecia.

Maxwell, M. (2002). Resources for Morphology Learning and Evaluation.
Linguistic Data Consortium. Universidad de Pensilvania, EE. UU.

Megerdoomian, K. (2008). Finite State Morphology: A tutorial.
Computing Research Lab. Nuevo México, EE. UU.

Moreno, J. C. (2000). Curso universitario de lingüística general. Tomo II: Semántica, pragmática, morfología y fonología. 2a edición corregida.
Editorial Síntesis. Madrid, España.

Narayanan, A. & Hashem, L. (2002). On Abstract Finite-State Morphology.
Departamento de Ciencia Computacional, Universidad de Exeter, R. U.

Nojoumian, P. (2011). Towards the Development of an Automatic Diacritizer for the Persian Orthography based on the Xerox Finite State Transducer.
Departamento de Lingüística, Facultad de Artes, Universidad de Ottawa. Canadá.

Oliver, A. (2004). Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat.
Programa de doctorado Ciencia Cognitiva y Lenguaje, Departamento de Lingüística General, Universidad de Barcelona. Cataluña.

Oyarzo, C. (2008). La lengua mapuche (mapudungu(n)) hablada en Chile: sus principales rasgos estructurales.
Informe final de Seminario de Grado. Universidad de Chile. Santiago, Chile.

Payne, T. (2006). Exploring Language Structure. A Student's Guide.
Imprenta de la Universidad de Cambridge, R. U.

Payne, T. (2007). Describing Morphosintax. A guide for field linguists.
Universidad de Oregon & Instituto de Lingüística de Verano. Imprenta de la Universidad de Cambridge, R. U.

Probst, K. & Lavie, A. (2011). A structurally diverse minimal corpus for eliciting structural mappings between languages.
Instituto de Tecnologías del Lenguaje, Universidad Carnegie Mellon. Pensilvania, EE. UU.

Ríos, A. (2010). Applying Finite-State Techniques to a Native American Language: Quechua.
Trabajo final de Máster. Instituto de Lingüística Computacional, Universidad de Zürich, Suiza.

Rolleri, N. (2006). Réplicas a Microsoft. Mapuches usan Internet para defender su lengua. Estudiosos de la lengua mapuche alrededor del mundo reclaman el derecho de los propios usuarios a definir la forma en que se debe escribir su futuro.
Diario El Sur, Centro de documentación mapuche. Disponible en: http://www.mapuche.info/mapuint/sur050817.html

Rumian, S. (2011). Cada vez más cerca del traductor Español – Che Süngun…
Disponible en: http://millalikan.blogspot.com.es/2011/02/cada-vez-mas-cerca-del-traductor.html

Sak, H.; Güngör, T. & Saraçlar, M. (2009). A Stochastic Finite-State Morphological Parser for Turkish.
Dept. de Ingeniería Computacional & Dept. de Ingeniería Eléctrica y Electrónica, Universidad Boğaziçi. Estambul, Turquía.

Salas, A. (1992). Lingüística mapuche. Guía bibliográfica.
Revista Andina, 2, 475-537. Disponible en
http://www.facso.uchile.cl/publicaciones/sitios/lenguas/mapuche/salas/guia1.htm

Salas, A. (2006). El mapuche o araucano: Fonología, gramática y antología de cuentos.
Centro de estudios públicos. Santiago, Chile.

Schone, P. & Jurafsky, D. (2000). Knowledge-Free Induction of Morphology Using Latent Semantic Analysis.
Universidad de Colorado, EE. UU. En: Proceedings of CoNLL-2000 and LLL-2000, páginas 67-72, Lisboa, Portugal.

Smeets, I. (2008). A Gramar of Mapuche.
Mouton de Gruytier. Berlín, Alemania. Nueva York, EE. UU.

Smrž, O. (2004). Kenneth R. Beesley and Lauri Karttunen Finite State Morphology.
Instituto de Lingüística Formal y Aplicada, Facultad de Matemáticas y Física, Universidad Charles, Praga, República Checa. Disponible en: http://ufal.mff.cuni.cz/publications/year2004/pbml-finite.-pdf

Torero, A. (2005). Idiomas de Los Andes: Lingüística e historia. 2a edición.
Editorial Horizonte. Lima, Perú.

Trosterud, T. (2009). A constraint grammar for Faroese.
Universidad de Tromsø, Noruega. En Nealt Proceedings Series Vol. 8, Proceedings of the NODALIDA 2009 workshop, Constraint Grammar and robust parsing. Odense, Dinamarca. Disponible en:
http://hdl.handle.net/10062/14180

Wicentowski, R. (2002). Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework.
Universidad Johns Hopkins. Maryland, EE. UU.

Zúñiga, F. (2003). Some notes on the Mapudungun evidential.
Manuscrito. Universidad de Leipzig, Alemania.

Zúñiga, F. (2006). Mapudungun: El habla mapuche.
Centro de estudios públicos. Santiago, Chile.

Zúñiga, F. (2009a). An exploration of the diachrony of Mapudungun valency-changing operations. Universidad de Zürich, Suiza.

Zúñiga, F. (2009b). Review of A Grammar of Mapuche, by Ineke Smeets.
En International Journal of American Linguistics, 75(2):282-284. Zürich Open Repository and Archive. Universidad de Zürich. Disponible en: http://www.zora.uzh.ch

Zúñiga, F. & Herdeg, A. (2007). A closer look at Mapudungun inversion and differential object marking.
Universidad de Zürich, Suiza.