Analizador morfológico de la lengua quechua basado en software libre helsinkifinite-statetransducer (hfst)

Hugo David CalderonVilca, Flor Cagniy Cárdenas Mariñó, Edwin FredyMamani Calderon

hdcalderon@gmail.com, clavelyfcm@gmail.com, mcedwin@gmail.com

Universidad Nacional Micaela
Bastidas de Apurímac, Perú
Av. Arenas Nº 121 Abancay-Perú
Universidad Nacional del Altiplano
Puno, Perú
Av. Sesquecentenario Nº 1150 Puno-Perú

Resumen: *En esta investigación, se presenta la creación de un analizador morfológico para la lengua quechua basado en software libre Helsinki Finite-StateTransducer (HFST). La tecnología HFST un transductor de estado finito ha sido usado como analizador morfológico en los idiomas del lado europeo: Inglés, Finlandés, Francés, Alemán, Italiano, Sueco, Turco, entre otros. Por otro lado, quechua es una lengua aglutinante, diferente a las lenguas europeas, por lo que en esta investigación se experimenta la adaptación de la tecnología HFST como herramienta de análisis morfológico, la misma que será un módulo y parte del sistema de traducción automática entre español y quechua en la plataforma Apertium. Para el experimento, se elige el Quechua del Este de Apurímac (qve), creando un diccionario LEXC compatible con HFST, definiendo lexicones para cada categoría gramatical, lexicones para cada sufijo, insertando más de 2000 palabras, entre raíces, adjetivos, verbos, adverbios y otras categorías gramaticales. Se compila el diccionario monolingüe del quechua tanto para el analizador morfológico como para su generación de palabras. Finalmente, como resultado se tiene el analizador morfológico (qve), se realiza las pruebas con entradas de palabras aglutinadas, teniendo como salida la palabra raíz y una serie de <etiquetas> que representan categorías gramaticales de los sufijos que lo acompañan.*

Palabras clave: Análisis morfológico, quechua, HFST, traductor automático, software libre.

Abstract: *This research presents the creation of a morphological analyzer for the Quechua language based on free software Helsinki Finite-State Transducer (HFST). HFST Technology finite state transducer has been used as a morphological analyzer tool in the languages of Europe side: English, Finnish, French, German, Italian, Swedish, Turkish and others. On the other hand Quechua is an agglutinative language that is different from European languages, that is why, in this research it is experimented the adaptation on of the HFST technology, it will be a module and part automatic translation system between Spanish and Quechua on the platform Apertium. For the experiment it has been choosen the Apurimac Quechua East (qve), creating a dictionary LEXC with support HFST, defined lexicons for each grammatical category, lexicons for each suffix, inserting more than 2000 words between roots, adjectives, verbs, adverbs and other categories grammatical. Monolingual dictionary is compiled for morphological analyzer as for generating words, finally as resulted has had the morphological analyzer (qve), testing is performed with agglutinated word entries having as output the root word and a series of grammatical categories <marks> representing the suffixes that accompany it.*

Keywords: Morphological analysis.quechua, HFST, machine translator, free software.

1 Introducción

Ante la existencia de múltiples culturas e idiomas se han desarrollado traductores automáticos como aplicación del procesamiento de lenguaje natural que aportan significativamente en el mundo de la informática permitiendo al ser humano comprender e interrelacionarse con sus semejantes mediante la traducción de textos o habla de un lenguaje natural a otro. Sin embargo, dichos avances, como los traductores automáticos, poco trascienden todavía en las minoritarias como el quechua. Por lo que en el camino de desarrollo e implementación de un traductor automático entre español y quechua es imprescindible el subsistema analizador morfológico en este caso el de quechua. Por eso en esta investigación, se experimenta la implementación de una analizador morfológico para la lengua quechua basado en el software libre Helsinki Finite-StateTransducer, una máquina de estado finito y un conjunto de herramientas para aplicaciones de analizadores morfológicos inicialmente usada para idiomas del lado europeo, tales como: Inglés, Finlandés, Francés, Alemán, Italiano, Sueco, Turco, entre otros.

El aporte fundamental de este trabajo de investigación es averiguar si el sistema HFST es funcional como herramienta de análisis morfológico de la lengua quechua camino a construir traductores automáticos con la lengua quechua. Por ende, esto complementará a las investigaciones de la lingüística quechua, tanto para estudiantes, como para profesionales e investigadores, quienes tendrían la posibilidad de obtener el resultado del análisis morfológico de esta lengua. Asimismo, la investigación busca la reivindicación de esta lengua que cumple un papel trascendental como vehículo de expresión y pensamiento.

Durante el experimento, se implementa el diccionario morfológico de la lengua quechua en un fichero LEXC

compatible con HFST especificando las categorías gramaticales mediante LEXICONES y subLEXICONES para cada sufijo, seguidamente compilando el diccionario para el analizador morfológico y el de generador de palabras. Finalmente, se realizan pruebas con entradas de diferentes palabras aglutinadas de la lengua quechua que pertenecen a las diferentes categorías gramaticales obteniendo como resultado en la salida la palabra aglutinada divida en: palabra raíz y una serie de símbolos <etiquetas> que representan categorías gramaticales de los sufijos que le acompañan, cuya interpretación responde a los fundamentos lingüísticos de la lengua quechua.

En la presente investigación, para el experimento, se ha tomado el Quechua del Este de Apurímac "qve", código clasificado de acuerdo con el SIL Internacional ISO-639-3, y según la otra clasificación que pertenece al tipo QII-C.

En la sección 1, se presenta la introducción de la investigación; en la sección 2, se presenta los trabajos previos o antecedentes de la investigación y la teoría del dominio que es el marco referencial; en la sección 3, están los experimentos y los resultados de la investigación; finalmente, en la sección 4, se plantean las conclusiones y trabajos futuros de la investigación.

2 Teoría del dominio y trabajos previos

2.1 Trabajos previos

Según el estudio "Corrector ortográfico, una lengua aglutinante: Quechua" realizada por [Rios2011], fundamenta que los métodos de corrección de ortografía desarrollado para idiomas como el Inglés, por lo general dependen de una lista completa de formas de las palabras completas, el requisito, de que no pueden ser satisfechas por idiomas morfológicamente complejas. Como resultado describe la implementación de un corrector ortográfico con métodos de estados finitos para la lengua aglutinante quechua (ISO 639-3: que).

[RiosG2009], en su investigación describe las características de la lengua quechua y su anotación morfológica y sintáctica, y demuestra cómo se alinean la lengua quechua con el idioma español mediante su análisis morfológico.

[Rataj2005] habla acerca la influencia del quechua en el español andino, donde detecta fenómenos ajenos al español general y peculiar del español andino. Dice, se dan tanto en el plano fonético e incluso fonológico, como en la morfología y sintaxis. Más tarde Rataj implementó el traductor automático en la dirección quechua español, tomando el quechua de Cusco (quz), sistema en construcción.

[Vargas2012] construye un sistema de texto a voz para el quechua estableciendo una representación fonética para el alfabeto quechua.

2.2 Procesamiento de lenguaje natural (PLN)

Las aplicaciones de Procesamiento de Lenguaje Natural son: Síntesis del discurso, Análisis del lenguaje, comprensión del lenguaje, reconocimiento del habla, síntesis de voz, generación de lenguajes naturales, traducción automática, recuperación de la información, dictado automático. Teniendo múltiples aplicaciones el Procesamiento del Lenguaje Natural contempla elementos como: análisis morfológico, análisis sintáctico, análisis semántico y análisis pragmático [Nils2004].

2.3 Lingüística Computacional (LC)

Es una rama de la inteligencia artificial (IA). Si bien las opiniones entre los especialistas divergen, se asume que el principal objetivo de la LC es la investigación y sistematización de la capacidad lingüística entendida como una capacidad cognitiva fundamental. Sucintamente, la LC se orienta hacia el estudio del conocimiento lingüístico obtenido a partir de la aplicación de un conjunto de formalismos y técnicas de representación. Con ello, se pretende el procesamiento del lenguaje natural (LN) mediante un ordenador. No se trata de elaborar modelos que posean realidad sicológica, sino más bien de construir modelos que simulen los tipos de conocimiento y los procesos que intervienen en la habilidad de transmitir e interpretar información a través del LN. En otras palabras, simular un conocimiento inteligente. Desde este punto de vista, se atribuye una cierta 'racionalidad' al ordenador, aunque es, por supuesto, estrecha y artificial [Vidal1996].

2.4 Traductor automático

Es una aplicación de Procesamiento de Lenguaje Natural, también considerada como área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. El traductor automático debe analizar el texto original, interrelacionar con la situación referida y, como resultado, debe encontrar el texto correspondiente en el lenguaje destino [Rusell2004].

El diseño de un sistema de Traducción Automática combina elementos de diversas disciplinas, especialmente la lexicografía, la lingüística, la lingüística computacional (la parte que se encarga de la implementación de las descripciones lingüísticas en algoritmos) y la Inteligencia Artificial (la parte de ésta que se encarga de la Representación del Conocimiento).

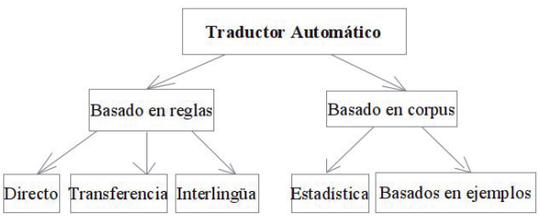


Figura 1: Modelos de Traducción Automática.

La traducción automática, basada en reglas, realiza transformaciones a partir del texto del idioma original reemplazando palabras por su equivalente en el idioma objetivo de traducción. En cambio la traducción basada en datos o corpus, realiza análisis de muestras reales en sus respectivas traducciones entre el par idiomas, mientras mayor cantidad de textos traducidos se tenga mejores resultados se obtiene.

2.5 Traducción Automática basada en reglas

La Traducción Automática basada en reglas establece tres enfoques principales: los enfoques directos, los de interlingua y los de transferencia (sintáctica y semántica).

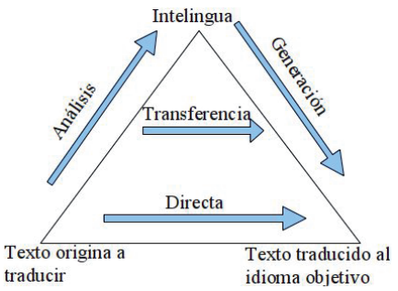


Figura 2: Paradigmas de TA basada en reglas.

Traducción automática por transferencia Modelo en la cual el texto original se analiza 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, modelo usado por la plataforma de código abierto Apertium.

2.6 Ingeniería de traducción de Apertium

[Armentano2007] Apertium es una plataforma de traducción automática de código abierto desarrollado por el grupo Transducens de la Universitatd'Alacant España, basado en reglas, cuya arquitectura usa transductores de estados finitos para el procesamiento léxico, modelos ocultos de Markov para la desambiguación léxica y procesamiento de patrones basado en estados finitos para la transferencia estructural, actualmente esta plataforma de traducción automática ha permitido implementar y poner en marcha a más de 35 pares de lenguas como sistemas de traducción automática, como se puede ver en [url:http://www.apertium.org]

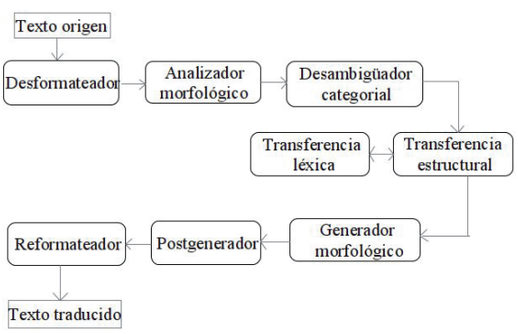


Figura 3: Arquitectura modular construido por la plataforma Apertium.

La plataforma Apertium proporciona: un ingenio de traducción independiente de las lenguas, herramientas para gestionar los datos lingüísticos necesarios para construir un sistema de traducción automática para un par de lenguas dado, la plataforma trabaja con los siguientes componentes: lttoolbox, apertium, apertium-lex-tools, OpenFST, Foma, HFST3 y vislcg3.

2.7 El analizador morfológico de Apertium

Segmenta el texto en formas superficiales (FS) (las unidades léxicas tal como se presentan en los textos) y entrega para cada FS una o más formas léxicas (FL) consistentes en un lema (forma base usada en los diccionarios clásicos), la categoría léxica (nombre, verbo, preposición, etc.) y la información de flexión morfológica (número, género, persona, tiempo, etc.). Las unidades léxicas de más de una palabra (multipalabras) son tratadas como formas léxicas individuales y, según su naturaleza, reciben un tratamiento específico, al recibir como entrada el texto del módulo anterior, el analizador morfológico proporciona como salida el texto resultante y con sus etiquetas que representan la categoría léxica ya sea nombre, verbo, adjetivo, etc, HFST es compatible con Apertium en cuanto al módulo de analizador morfológico.

2.8 Transductores de estados finitos

Es una derivación de los autómatas de estados finitos. La diferencia es que los transductores nos entregan como resultados un conjunto de símbolos que pertenecen al lenguaje. Un transductor también puede no producir ninguna salida para una cadena de entrada y en este caso se dice que el transductor rechaza la entrada. En general, los transductores de estados finitos son utilizados en análisis morfológico, en la investigación y aplicaciones de procesamiento del lenguaje natural. Todo transductor de estados finitos se puede convertir siempre en un transductor de letras [Forcada2012]. Un transductor de letras es una máquina idealizada que se compone de:

i) Un conjunto finito de estados: un único estado inicial donde el estado en el que se encuentra el transductor antes de procesar la primera letra o el primer símbolo de la entrada; uno o más estados de aceptación, a los que sólo se llega después de haber leído completamente una entrada válida y, por tanto, sirven para detectar las palabras válidas.

ii) Un conjunto también finito de transiciones de estado, cada una de las cuales se compone de: a) el estado de partida, b) el estado de llegada, c) la letra o símbolo de entrada, d) la letra o el símbolo de salida.

El transductor construye cada vez que lee un símbolo de la entrada una lista de estados vivos o activos, cada uno de los cuales tiene una salida asociada, una secuencia de símbolos. El funcionamiento de transductor de letras es distinto para cada tipo de operación de procesamiento léxico [Forcada2012].

2.9 Morfología lingüística

Consiste en determinar la forma, clase o categoría gramatical de cada palabra que hace parte de una oración, haciendo lo que se conoce como etiquetado morfológico [Mooney2003]. La morfología estudia los morfemas y sus

combinaciones en el interior de las palabras. Las palabras pueden estar compuestas por lexemas y morfemas: Lexema o raíz: es la parte de la palabra que no varía. Los lexemas son aquellas partes de las palabras que forman las unidades mínimas que contienen el significado de las palabras. Morfema: es la parte de la palabra que varía, es decir, la parte que se añade al lexema para completar su significado. Como para formar palabras nuevas y para completar su significado, pueden ser los accidentes del vocablo género, número, etc.

2.10 Lengua quechua

Quechua, también denominada Quichua es una familia de lenguas originaria de los Andes Centrales que se extiende por la parte occidental de Sudamérica. Es una macrolengua con una población hablante de más de 9'000,000 distribuidos en los países: Perú, Argentina, Ecuador, Chile y Bolivia, es lengua co-oficial en Perú.

Quechua, como macrolengua, se clasifica en 44 lenguas diferentes con código propio de acuerdo con el ISO 639-3 dada por SIL International, de los cuales 32 variantes de quechua se encuentran relacionadas con Perú, uno de ellos, tomado para el experimento (qve) Quechua del Este de Apurímac.

2.11 Helsinki Finite-State Transducer (HFST)

Es un transductor de estado finito destinado a la aplicación de analizadores morfológicos, y otras herramientas que se basan en la tecnología de transductores de estados finitos ponderados y no ponderados. HFST trabaja bajo la licencia GNU Lesser General PublicLicense v.3.0k.

2.12 LEXC

LEXC es un lenguaje de programación declarativo de alto nivel para especificar las reglas que permitan la combinación de lexemas y morfemas. Dichas reglas se especifican mediante los Lexicones que están asociados al compilador para definir el autómata de estado finito y los transducers, especialmente adecuado para la definición de los léxicos del lenguaje natural en un analizador morfológico. El resultado de la compilación de LEXC es un autómata de estado finito compatible con HFST. El LexiconRoot es el principal como marca de inicio del estado. En general, allí puede estar varios Lexicones, pero exactamente uno es el Root.

3 Experimentos y Resultados

3.1 Construcción del diccionario morfológico de la lengua quechua mediante LEXC

El diccionario morfológico se escribe en un archivo LEXC que permite al compilador definir el Autómata de Estado Finito Transducer. En el fichero LEXC, se ha escrito los lexicones para cada categoría gramatical definiendo reglas de la estructura de la palabra. Así mismo, se ha definido nombres, adjetivos, verbos, etc.

Primero, se ha definido los tipos de palabras en la cabecera del archivo LEXC; seguidamente, se definió LEXICON Root, que es la marca donde inicia el estado de la máquina, donde están definidos las categorías gramaticales: Nombres, Nombres Género, Nombres Derivados, Nombres Propios, Topónimos, Nombres Hispanos, Adjetivos, Adjetivos Género, Adjetivos Derivados, Pronombres, Pronombres Personales, Preadjetivos, Adverbios, Adverbios Temporales, Numerales, Posposiciones, Conjunciones, Interjeciones, Verbos, Verbos Transitivos, Verbos Intransitivos, Verbos de Movimiento, Puntuación, etc.

Luego se define otros LEXICONES para cada categoría gramatical mencionada, así mismo un LEXICON para cada sufijo, LEXICON CLIT-Final, LEXICON CLIT-chu, LEXICON CLIT-taq, LEXICON CLIT-pas, LEXICON CLIT-raq-na, LEXICON CLIT-puni, LEXICON CLIT-lla, LEXICON N-POSTcaso-wan, LEXICON N-POSTcaso, LEXICON N-lla-POSTcaso, LEXICON N-FLEX-Caso-lla, LEXICON N-FLEX-lla-Caso, LEXICON N-FLEX-Caso, LEXICON ADJ-DER, LEXICON NUM-Poss, LEXICON V-PERS, LEXICON V-FUT.

LEXICONES de clase de palabra. Finalmente se define un LEXICON para insertar las las palabras: LEXICON N, LEXICON N-M, LEXICON N-F, LEXICON NP-ANT-M, LEXICON NP-TOP, LEXICON ADJ, LEXICON PRON, LEXICON ADV.

La estructura del diccionario monolingüe de la lengua quechua (QVE) consiste en un conjunto de sentencias o bloques que se definen para el LEXICON Root y LEXICONES para cada categoría gramatical y LEXICONES para cada tipo de sufijos. Este conjunto de instrucciones está definido en el archivo apertium-es-qve.qve.lexc compatible con HFST.

Cada entrada proporcionada al sistema debe dar salidas con la palabra raíz y agregando etiquetas de los sufijos y palabras compuestas, de manera que cada etiqueta tiene su significado que pertenece a las categorías gramaticales o el tipo de sufijo, por tanto, cada palabra, sea raíz o sufijo, se ajustan a las siguientes tablas:

Tabla 1: Tipos de palabras mediante símbolos.

Símbolos	Significado
%<n%> %<prn%> %<adj%> %<adv%> %<vblex%> %<m%> %<f%> etc.	! nombre ! pronombre ! adjetivo ! adverbio ! verbo ! masculino ! femenino

En la Tabla 1 muestra la definición de los tipos de palabras de la lengua quechua en la cabecera del archivo LEXC, dicha definición se implementa mediante la asignación de los símbolos que representan en adelante para referirse desde los LEXICONES, el símbolo “%” es para el escape de la secuencia, su ausencia no permitiría como entrada de los otros símbolo “< y >”.

Tabla 2: Especificación del LEXICON Root.

LEXICON Root
Nombres ; Pronombres ; Adjetivos ; Adverbios ; Verbos ; etc

En la Tabla 2, se lista todas las categorías gramaticales de la lengua quechua, las mismas que se enlazarán con cada palabra insertada definidas mediante los LEXICONES dadas desde la Tabla 07.

Tabla 3: Definición de LEXICON – clíticos.

LEXICON CLIT-Final! sufijos finalistas que no aceptan más sufijos
%<topi%>:%>qa # ; %<rnd%>:%>ri # ; %<rep%>:%>sis # ; %<cnj%>:%>chá # ; %<dub%>:%>chus # ; %<dub%>:%>chusuna # ; %<dub%>:%>chusina # ;

En la Tabla 3, se define los clíticos. Son sufijos finalistas que no aceptan otros sufijos después de su aplicación. Los símbolos Dir/LR indican la restricción que sólo se analiza la palabra, pero no se genera. En cambio, Dir/RL es la restricción que no se analiza la palabra pero si genera.

Tabla 4: Definición de LEXICON sufijos.

LEXICONES ! sufijos que aceptan otros sufijos
LEXICON CLIT-chu %<qst%>:%>chu # ; ! -chu-qa(ya), -chu-ri CLIT-Final ; LEXICON CLIT-taq %<contr%>:%>taq CLIT-chu ; %<contr%>%<excl%>:%>táq # ; CLIT-chu ;

En la Tabla 4, se implementa, en el diccionario, el resto de los sufijos, es decir, todos los sufijos. La diferencia con laTabla 3 es que estos sufijos son de todo tipo y aceptan otros sufijos después de su aplicación. De esta manera, las palabras generadas de la lengua quechua son aglutinantes.

Tabla 5: LEXICON casos.

LEXICONES
LEXICON N-POSTcaso-kama ! kama - <ter>×<distrib> ! kama-lla CLIT-puni ; LEXICON N-POSTcaso-wan %<cnjcoo%>:%>wan CLIT-puni ; %<cnjcoo%>:%>puwan CLIT-puni ; %<cnjcoo%>:%>piwan CLIT-puni ; CLIT-puni ;

En la Tabla 5, se definen los casos de la lengua quechua: son sufijos que permiten generar las palabras aglutinadas.

Tabla 6: LEXICONES de categorías gramaticales.

LEXICON N %<n%>: N-DER ; ! %<n%>%<GD%>: N-DER ; LEXICON N-M %<n%>%<m%>: N-DER ; LEXICON N-F %<n%>%<f%>: N-DER ; LEXICON PRON %<prn%>: N-DER ; LEXICON ADV %<adv%>: CLIT-puni ; LEXICON ADV-lla %<adv%>: CLIT-lla ; LEXICON ADV-taq %<adv%>: CLIT-taq ; LEXICON V-IV %<vblex%>%<iv%>: V-DER ; etc.

En la Tabla 6, se definen la combinación de las diferentes categorías gramaticales expresadas como reglas de la morfología.

Tabla 7: LEXICON nombres.

Nombres	Significado
%wawa:wawa% N;	!niño (a)
%alqu:alqu% N ;	! "perro"
%amawta:amawta% N ;	! maestro
%asnu:asnu% N ;	! "asno, burro"
%wasi:wasi% N ;	! casa

En la Tabla 7, se implementa los nombres de la lengua quechua con su indicador “N”. símbolo que representa la pertenencia a la categoría gramatical nombres.

Tabla 8: LEXICON adjetivos

Nombres	Significado
%ashka%askha%ADJ;	!“harto
%mishk'i% ADJ;	!“dulce”
%maqlla% ADJ;	!“tacaño”
%unqusqallaña% ADJ;	!“enfermo”
%manchasqallaña%ADJ;	!“asustado”

En la Tabla 8, se implementa los adjetivos de la lengua quechua con su indicador “ADJ”, símbolo que representa la pertenencia a la categoría gramatical adjetivos.

Tabla 9: LEXICON adverbios.

Adverbios	Significado
%apuraylla% ADV;	!“rápido”
%p'itaylla% ADV;	!“corriendo”
%allinta% ADV;	!“bien”
%mana%allintachu%ADV;	!“mal”
%sumaqchata% ADV;	!“bonito”

En la Tabla 9, se implementa los adverbios de la lengua quechua con su indicador “ADV”, símbolo que representa la pertenencia a la categoría gramatical adverbios.

Tabla 10: LEXICON verbos de movimiento.

Verbos		Significado
kutiy:kuti	V-VM;	!regresar
khuyay:khuya	V-TV;	!querer, apreciar
k'iriy:k'iri	V-IV;	!herir
kusikuy:kusiku	V-IV;	!alegrarse
llakiy:llaki	V-IV;	!apenarse,

En la Tabla 10, se implementa uno de los tipos de verbos de la lengua quechua con su indicador “V-VM”, “V-TV” y “V-IV”, símbolos que representan la pertenencia a la categoría gramatical de diferentes verbos.

3.2 Compilación del diccionario morfológico

HFST cuenta con herramientas para su compilación, ejecución y más manipulación de todo lo concerniente al análisis morfológico las siguientes instrucciones fueron necesarias en la compilación:

Compilando diccionario morfológico quechua

#cat apertium-es-qve.qve.lexc | grep -v 'Dir/RL' > .deps/qve.LR.lexc

Esta instrucción crea un archivo nuevo apartir de LEXC quitando los que están marcados con Dir/RL (dirección Right Left). Teniendo como resultado las palabras que están marcadas se deben analizar, pero no generarse:

1) nishu:nishuPreadj ; ! "demasiado"

2) nishu:nishiwPreadj ; ! "demasiado" - Dir/LR

3) nishu:nisiwPreadj ; ! "demasiado" - Dir/LR

4) nishu:nisyuPreadj ; ! "demasiado" - Dir/LR

Mostrando una parte del archivo LEXC 1), la palabra se analiza y se genera. El análisis es utilizado cuando sea la traducción en la dirección de quechua a español, en cambio las palabras 2), 3) y 4) tienen la marca Dir/LR que significa que se analizan pero no se deben generar, es decir, en cuanto la traducción sea en la dirección español a quechua se utiliza el generador, por tanto, al llevar esta marca dichas palabras, no pueden ser generados de manera que la única forma que la palabra “demasiado” sea traducido a quechua es “nishu”.

#hfst-lexc --formatfoma .deps/qve.LR.lexc -o .deps/qve.LR.lexc.hfst

Esta instrucción pone en formato foma el archivo LEXC, forma es una librería compatible con HFST.

#hfst-twolc --formatfomaapertium-es-qve.qve.twol -o .deps/qve.twol.hfst

Instrucción que pone en formato foma el fichero twol donde se indican algunas reglas de la morfología.

#hfst-compose-intersect -1 .deps/qve.LR.lexc.hfst -2 .deps/qve.twol.hfst -o .deps/qve.LR.

Proceso que compone el lexc y twol en solo fichero.

#hfsthfst-invert .deps/qve.LR.hfst | hfst-fst2fst -O -o qve-es.automorf.hfst

Proceso final que da como resultado el diccionario de análisis morfológico de la lengua quechua cuyo fichero resultante es qve-es.automorf.hfst.

Compilando diccionario de auto generación

#cat apertium-es-qve.qve.lexc | grep -v 'Dir/LR' > .deps/qve.RL.lexc.

Esta instrucción crea un archivo nuevo apartir de LEXC quitando los que están marcados con Dir/LR (dirección Left Right). Teniendo como resultado las palabras que están marcadas no se analizan, pero si deben generarse.

#hfst-lexc --formatfoma .deps/qve.RL.lexc -o .deps/qve.RL.lexc.hfst.

Esta instrucción que pone en formato foma una librería compatible con HFST.

#hfst-twolc --formatfomaapertium-es-qve.qve.twol -o .deps/qve.twol.hfst

Instrucción que pone en formato foma el fichero twol donde se indican algunas reglas de la morfología.

#hfst-compose-intersect -1 .deps/qve.RL.lexc.hfst -2 .deps/qve.twol.hfst -o .deps/qve.RL.

Proceso que compone el lexc y twol en solo fichero.

#hfst hfst-fst2fst -O .deps/qve.RL.hfst -o es-qve.autogen.hfst.

Proceso final que da como resultado el diccionario de autogeneración de palabras utilizado para el traductor en la dirección español a quechua, cuyo fichero resultante es qve-es.autogen.hfst.

3.3 Resultados de la construcción del diccionario de la lengua quechua

El diccionario monolingüe de la lengua quechua del Este de Apurímac está implementado en el fichero apertium-es-qve.qve.lexc, con más de 2000 palabras que son raíces especificado cada palabra con su categoría gramatical a la que pertenece. Su compilación ha permitido crear dos ficheros: el fichero binario qve-es.automorf.hfst que sirve para el analizador morfológico y el fichero binario qve-es.autogen.bin para la generación de palabras.

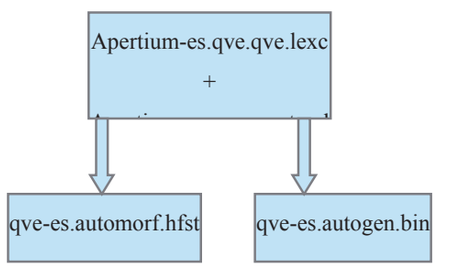


Figura 4: Diccionario monolingüe de quechua compilado.

Según [Armentano2007] y según [Forcada2012], el diccionario monolingüe está especificado en formato XML en los ficheros .dix y no LEXC, esta investigación experimenta la implementación del diccionario

monolingüe compatible con HFST, desde su creación hasta su compilación es diferente; sin embargo una vez compilado el diccionario construido en esta investigación es compatible con la plataforma de código abierto Apertium, de manera que el analizador morfológico presentado en esta investigación, sería parte del sistema de Traducción Automática de par de lenguas español y quechua basado en la plataforma de traducción automática Apertium.

Descripción de los ficheros principales del analizador morfológico y ficheros relacionados:

apertium-es-qve.qve.lexc. Diccionario monolingüe de quechua compatible con HFST.

apertium-es-qve.qve.twol. Reglas de morfología compatible con HFST.

La prueba de funcionamiento del analizador morfológico es una secuencia de comandos que debe aceptar la herramienta HFST:

#echo “wasiykuna” | hfst-proc -x qve-es.automorf.hfst

Teniendo como resultado: wasi<n><px1sg><pl><nom> la cual indica que “wasi” es la palabra raíz, <n> indica que pertenece a la categoría nombres, <px1sg> primera persona singular, <pl> plural <nom> nominativo.

Aplicando la generación de la palabra se tiene:

#echo “wasi<n><px1sg><pl><nom>” | hfst-proc $1 es-qve.autogen.hfst

Se tiene como resultado: wasikuna, explicando que al detectar la raíz wasi y con sus respectivas etiquetas genera la palabra “wasikuna”.

Además se observa que la ejecución es invocando a HFST-PROC con sus parámetros correspondientes, otra vez según Armentano, *et al*, (2007) y según Forcada, Vvoney y Oortiz (2012), la ejecución sería invocando LT-PROC, otra herramienta de analizador morfológico utilizado por la plataforma de traducción automática de código abierto Apertium.

3.4 Resultados delanálisis morfológico de palabras

#echo "wasiykuna" | hfst-proc -x qve-es.automorf.hfst

Entrada nombre y plural: wasiykuna

Resultado: wasi<n><px1sg><pl><nom>

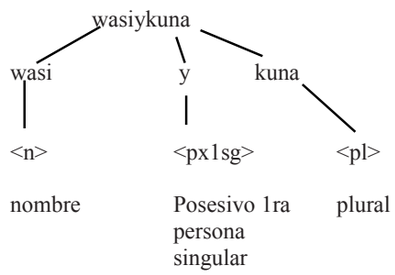


Figura 5: Análisis morfológico de la palabra **wasiykuna.**

#echo "sumaqchallata" | hfst-proc -x qve-es.automorf.hfst

Entrada adjetivo nombre y sufijo: **sumaqchallata**

Resultado: sumaq<adj><dim><lim><acc>

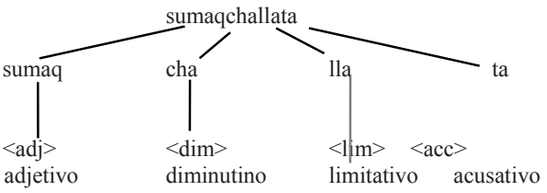
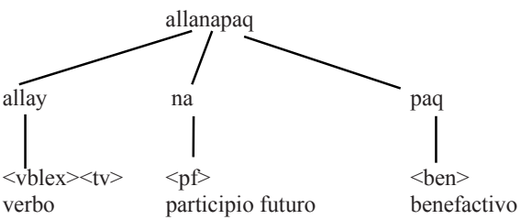


Figura 6: Análisis morfológico de la palabra sumaqchallata.

echo "allanapaq" | hfst-proc -x qve-es.automorf.hfst

Entrada verbo y sufijo: allanapaq

Resultado: allay<vblex><tv><pf><ben>



verbo transitivo

Figura 7: Análisis morfológico de la palabra allanapaq.

echo "maypiraq" | hfst-proc -x qve-es.automorf.hfst

Entrada adverbio y sufijos: maypiraq

Resultado: may<adv><itg><loc><cnti>

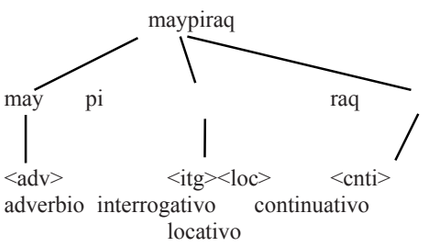


Figura 8: Análisis morfológico de la palabra **maypiraq.**

4 Conclusiones y trabajos futuros

4.1 Conclusiones

Se presenta el analizador morfológico de la lengua quechua basado en el software libre Helsinki Finite-StateTransducer (HFST) con el diccionario monolingüe de la lengua quechua implementado y compilado con más de 2000 palabras entre sus diferentes categorías gramaticales en formato LEXC compatible con HFST. Contrastando con [Armentano2007] y según

[Forcada2012], el diccionario monolingüe está especificado en formato XML en los ficheros .dix en contraste. En esta investigación, se experimenta la implementación del diccionario monolingüe compatible con HFST en formato LEXC. Cabe señalar que una vez compilado es similar al de formato .dix y cumple lo mismo como módulo de analizador morfológico para a plataforma de traducción automática de código abierto Apertium. La tecnología Helsinki Finite-State Transducer es funcional, utilizable y se adapta para el analizador morfológico de las lenguas aglutinantes como es el quechua. Finalmente, todos los datos lingüísticos implementados tienen la característica de software libre y pueden ser usados para cualquier propósito.

4.2 Trabajos futuros

Helsinki Finite-State Transducer también sería aplicable para la lengua Aymara sabiendo que con quechua son de la misma familia de las lenguas aglutinantes. El analizador morfológico de quechua es utilizable como módulo como parte de Traductor Automático español-quechua basado en el sistema de código abierto Apertium que se quiera construir. Sería una investigación de mayor para escribir un diccionario morfológico que acepte todas las variantes de la macro lengua quechua.

Referencias bibliográficas

[Armentano2007] Armentano-Oller Carme, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Marco A. Montava Belda, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez y Felipe Sánchez-Martínez. Group Transducens Department de Llenguatges i Sistemes Informàtics Universidad d'Alacant: Apertium una plataforma de código abierto para el desarrollo de sistemas de traducción automática, 2007.

[Breña2003] Breña Ramón: Autómatas y Lenguajes. Tecnológico de Monterrey, Campus Monterrey, México. 2003.

[Cerron1987] Cerron Palomino Rodolfo: Lingüística Quechua–Centro de Estudios Rurales Andinos Bartolomé de las Casas, 1987.

[Diaz2007] Díaz de Ilarraza, A. Mayor and K. Sarasola: Reutilización de recursos lingüísticos en la construcción de un sistema de Traductor Automático inglés-euskara, 2007.

[Forcada2012] Forcada Mikel L., Boyan Ivanov Vonvev, Sergio Ortiz Rojas, y otros: Documentación del sistema de código abierto OpentradApertium de Traducción Automática de transferencia sintáctica superficial. Departament de Llenguatges i Sistemas Informàtics Universitat d' Alacant. 2012.

[Kenneth2002] Kenneth R. Beesly y LauriKarttunen.Finite-State MorphologyXerox Tools and Techniques. 2002.

https://victorio.uit.no/langtech/tags/Root-of-gt-UTF-8-conversion/gt/doc/book.pdf_1.pdf

[Mooney2003] Mooney y Raymond J.: Oxford Handbook of Computacional Linguistics. Oxford University Pres, 2003.

[Rataj2005] Rataj Vlastimil: La Influencia del Quechua en el Español Andino.2005.

[Rios2012] Rios Annete y Wolk Martin: Parallel Treebanking Spanish-Quechua, Linguistic Issues in Language Technology – LiLT, 2012.

[Rios2009] Rios Annete y WolkMartin., A Quechua-Spanish Parallel Treebank. University of Zurich - Zurich Open Repository and Archive, 2009.

[Rios2011] Rios Annete: Spell Checking an Agglutinative Language: Quechua. University of Zurich - Zurich Open Repository and Archive, 2011.

[RiosG2009] Rios Annete, Gohring A. y Wolk M.: A Quechua-Spanish Parallel Treebank. University of Zurich – Zurich Open Repository and Archive, 2009.

[Rusell2004] Rusell, Stuart y Norvig, Meter: Inteligencia Artificial un enfoque moderno. Segunda Edición. Madrid. Pearson Educación S.A.2004.

[UANCV2004] Universidad Andina Néstor Cáceres Velásquez-Perú: Morfología Contrastiva Quechua/Aymara/Castellano. Escuela de Postgrado de la, Segunda Especialización en Educación Bilingüe Intercultural, 2004.

[Vargas2011] Vargas John, Cruz Juan A and Richard Castro Richard: Let's Speak Quechua The Implementation of aText-to-Speech System for the Incas' Language2011.