UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMATICA

TESIS DE GRADO

"DISEÑO DE UN AUTÓMATA PARA LA TRADUCCIÓN DEL IDIOMA CHINO AL ESPAÑOL"

PARA OPTAR AL TÍTULO DE LICENCIATURA EN INFORMATICA
MENCION: INGENIERIA DE SISTEMAS INFORMATICOS

POSTULANTE: BLANCA REINA CHOQUE GARCÍA
TUTOR METODOLOGICO: LIC. JAVIER REYES PACHECO
ASESOR: LIC. GERMAN HUANCA TICONA

LA PAZ – BOLIVIA
2015

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

LA CARRERA DE INFORMÁTICA DE LA FACULTAD DE CIENCIAS PURAS Y NATURALES PERTENECIENTE A LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.

LICENCIA DE USO

El usuario está autorizado a:

a) visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
b) copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
c) copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la referencia correspondiente respetando normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

Dedicado a mi mamá, porque sin ti nada de esto hubiera sido posible. Gracias por enseñarme a seguir adelante a pesar de todo.

AGRADECIMIENTO

Primeramente quiero agradecer a Dios, por estar a mi lado siempre, tanto en los momentos tristes y difíciles, como también en los alegres; gracias por cuidarme y protegerme cuando llegaba tarde a casa por cuestiones de estudio ó de trabajo. Gracias padre mío.

Agradezco a mi mamá Ambrocia García, por brindarme siempre su apoyo moral y confiar en mí, por enseñarme a luchar en la vida y hacerme fuerte a pesar de las malas circunstancias de la vida.

Agradezco a mis hermanos Fanny, Jhonny y Odilia por sus consejos, por sus palabras de ánimo y corrección. Gracias por apoyarme y comprenderme.

Gracias……a mis amigas del alma Lidia, Deysi y Verónica; gracias por su sincera amistad. Con ustedes la vida universitaria fue mucho más llevadera, las tendré siempre en mi corazón.

También deseo expresar mi gratitud al Lic. Javier Reyes Pacheco y al Lic. Germán Huanca, por su tiempo, por sus enseñanzas, consejos y su paciencia en la elaboración del presente trabajo.

Por último agradecer a los bibliotecarios de la carrera de Informática a Dn. Fernando y a Dn. Daniel gracias por colaborarme en la selección de libros a lo largo de mi carrera universitaria.

RESUMEN

Hoy en día las brechas de distancia y tiempo ya no existen gracias a la tecnología; se puede enviar rápidamente un correo (e-mail) ó estar presente en una reunión muy importante por teleconferencia.

Así también la tecnología debe colaborarnos en el aprendizaje de nuevos idiomas, sobre todo los más difíciles y contribuir de esta manera a un aprendizaje mucho mas integro.

Es pues el chino mandarín el idioma más hablado en el mundo. En la república popular de China existen 1300 millones de habitantes y más del 95% lo habla.

Es también el idioma hablado en Taiwán, Hong Kong, Singapur, en Malasia y naturalmente en todas las comunidades chinas establecidas en los cinco continentes.

El Hanyu ("lengua del pueblo Han") es la lengua oficial en China donde también se la llama **pǔ tōng huà** (la lengua corriente), lo que en español llamamos chino mandarín.

Con la ayuda de la teoría informática probada y fundamentada, para el reconocimiento de lenguajes formales como son los **Autómatas finitos deterministicos** se hizo un estudio, que nos permita la traducción de los caracteres chinos a través de su correspondiente pīnyīn. Haciendo uso de software OCR para el reconocimiento de dichos caracteres.

Los autómatas son modelos informáticos útiles; analizadores léxicos de un compilador (identifican y clasifican las palabras de un lenguaje: identificando literales y operadores).

Lo descrito anteriormente es pues a groso modo una breve resumen de la presente Tesis.

CONTENIDO

2. MARCO TEORICO

INDICE DE FIGURAS

1. PRESENTACIÓN

1.1 INTRODUCCIÓN

Se ha visto que las naciones mantienen intercambios en telecomunicaciones, comercio internacional, investigación científica, becas de estudio, etc. Lo cual obliga que el idioma no sea un impedimento, para el desarrollo de dichas actividades.

Solo en nuestro país, las importaciones que tenemos con el país de China han alcanzado un 75% .Y que decir en Telecomunicaciones, que el personal que opera el satélite Túpac Katari tuvo que ir a capacitarse a China.

De esta manera es que surge la propuesta de una investigación en este tema del **traductor de caracteres chinos utilizado autómatas**, por medio reconocedor óptico de caracteres (OCR).

Es por tal motivo que manifiesta importancia tanto para los que estudian el idioma, como para los que circunstancialmente lo han visto en el trabajo ó en el comercio con aquel país.

Los autómatas son modelos informáticos útiles; analizadores léxicos de un compilador (identifican y clasifican las palabras de un lenguaje: identificando

literales y operadores). Utilizados en software sirven para explorar grandes corpus de texto (colección de páginas web), en busca de palabras, frases, etc. El reconocimiento óptico de caracteres (OCR), la digitalización de texto, son aplicaciones informáticas. Estos identifican automáticamente símbolos ó caracteres que pertenecen a un determinado alfabeto a partir de una imagen base para guardarlos en un formato compatible con programas de edición de texto.

Los algoritmos de Reconocimiento Óptico de Caracteres tienen la finalidad de diferenciar un texto, de una imagen cualquiera. Para hacerlo se basan en 4 atapas bien definidas: preproceso que es la binarización, fragmentación ó segmentación de la imagen, representación de la imagen (extracción de características) y reconocimiento (comparación con patrones).

El buen funcionamiento de la mayoría de los programas OCR se basa en la buena definición de esta última etapa. Existen diferentes métodos para llevar a cabo la comparación.

Comparación con patrones se trata de una de las etapas más importantes, se compara todos los caracteres obtenidos, con los caracteres de la aplicación guardados anteriormente en una base de datos del mismo.
Sintáctico estructural estudia la estructura de los objetos, es decir, usa teoría de lenguajes formales, gramáticas, teoría de autómatas, etc.
En la escritura china uno de los rasgos más resaltantes es que el carácter suele coincidir con una sílaba que posee significado.
La mayor parte del léxico chino moderno se compone de palabras monosílabas, un carácter es pues una palabra, entendiendo como palabra, una unidad léxica que se puede combinar libremente en una frase.

También existe una gran cantidad de palabras compuestas de dos o varios caracteres que pierden entonces un poco de su valor semántico original para

formar un nuevo conjunto independiente con el ó los caracteres que se le añaden.

Existen diversos criterios para clasificar los tipos de caracteres chinos. Lo más sencillo es dividirlos en tres categorías básicas:

Los caracteres más antiguos son los pictogramas, esto es, dibujos del concepto que representan.

El segundo tipo de caracteres son los llamados ideogramas. En estos casos los pictogramas se combinan para sugerir ideas por asociación.

El tercer tipo de caracteres lo constituyen los logogramas. Este tipo abarca la inmensa mayoría de los caracteres chinos actuales; consiste en la modificación de otro carácter con el que comparte pronunciación añadiéndole otro componente que lo distingue. El componente añadido es un radical que aporta una idea semántica respecto al tipo de significado representado por el nuevo carácter.

En lo que concierne a la transcripción fonética del chino, existe hoy en día una transcripción universal, el **Hànyǔ pīnyīn**, literalmente “deletrear la lengua de Han” que permite notablemente a las personas de otros países aprender la lengua china. Esta transcripción preconizada por el gobierno Chino, se enseña también en los colegios con el mismo propósito: pronunciar correctamente el idioma nacional, por medio de las letras latinas, funciona como un sistema cerrado, un poco como el alfabeto fonético internacional.
Pero ello implica un aprendizaje especifico, para saber por ejemplo que **qù** se pronuncia “ch’u”, **zchou** se pronuncia “d-cho-u”.

El **pīnyīn** permite transcribir en letras latinas los sonidos de los caracteres chinos, y permite una pronunciación más o menos correcta al no iniciado y ser entendido en tierra China.

1.2 ANTECEDENTES

La actual investigación cobra interés respecto a una de las más enigmáticas culturas y una de las más antiguas del mundo.

En la Universidad Mayor de San Andrés en la carrera de Informática, se estudiaron y desarrollaron trabajos similares pero con respecto al español, aymara y texto braille, que se detallaran a continuación:

- Condori, “Traductor del idioma español – aymara utilizando deducción natural”, su objetivo realizar un modelo textual y de retroalimentación por relevancia para una traducción natural por inducción del español al aymara, 2007.
- Mercado, ”Reconocimiento y traducción de texto Braille a idioma nativo”, su objetivo principal es desarrollar un método que permita el reconocimiento del braille y posteriormente traducirlo al lenguaje aymara,2007
- Paye, “Traducción semántica y sintáctica del español al aymara usando autómatas”, su objetivo diseñar un prototipo el cual sea capaz de traducir la lengua del español en una forma gramatical correcta al aymara, 2012.

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 PLANTEAMIENTO DEL PROBLEMA

Tabla 1.1: **Causa – Efecto**

PROBLEMA	CAUSA	EFECTO	SOLUCIONES
La no comprensión de la escritura china y el no tener una comunicación fluida con el estado Chino	La falta de libros, diccionarios, software, recursos pedagógicos que permitan aprender el idioma.	La perdida de oportunidades de trabajos, becas, realización de mejores negocios, etc.	El estudio factible de métodos informáticos que permita la traducción de los caracteres chinos.
El estudio del significado de las palabras en chino fue ajeno a los estudios lingüístico-gramaticales sobre todo al que se refiere a los cambios de significado.	Se creía que dicho problema atañía más a la filosofía (lógica) que a los estudios estrictamente lingüísticos.	El desconocimiento de la gramática y léxico de la lengua china.	La significación de una palabra, es su uso en la lengua (lo primero que se percibe son las diferencias significativas entre palabras que aparecen juntas)
El desconocimiento de la relación entre significado y descripción lingüística.	Si el lenguaje era natural ó artificial (resultado de una convención); si el signo lingüístico es o no arbitrario.	Preguntarse ¿si se adjudican los nombres a las cosas por su naturaleza o por un pacto social?	El Estado Chino promovió el sistema pīnyīn en 1948 y en 1980 se generalizó, este permite una correcta pronunciación y aprendizaje de la lengua china.

1.3.2 FORMULACIÓN DEL PROBLEMA

¿El desarrollo de una investigación sobre un traductor, que utiliza autómatas formales, a partir de un Reconocedor Óptico de Caracteres (OCR), serán adecuados para la traducción de caracteres chinos?

1.3.3 DELIMITACIÓN DEL PROBLEMA

El traductor tendrá base en el (chino simplificado: 汉字, pinyin: hànzì) que se emplean en la escritura del idioma chino y no así en el chino tradicional: 漢字, además para el prototipo solo se tomara en cuenta 3.000 caracteres que se utilizan en la lengua común, de los 10.000 que son utilizados en su totalidad de la lengua china. Por último, se reconocerá texto tipográfico y no así manuscrito.

1.4 HIPÓTESIS

La aplicación de un modelo de autómata que permita la traducción del idioma chino al español haciendo uso del pīnyīn.

1.5 OBJETIVOS

1.5.1 OBJETIVO GENERAL

Diseñar un Autómata para la traducción de caracteres chinos mediante su correspondiente* pīnyīn*.

1.5.2 OBJETIVO ESPECIFICO

1. Desarrollar un prototipo del traductor utilizando Visual.Net. y SQL 2008
2. Implementar el reconocimiento óptico de caracteres (OCR) para el procesamiento de la imagen off-line, que trabaje conjuntamente con el sistema basado en autómatas semánticos para posterior comparación con los logogramas de la base de datos del traductor.
3. Ante el reconocimiento de un nuevo logograma, permitir la actualización constante de la base de datos.
4. Analizar la fase de reconocimiento más factible en el OCR.
5. Evaluación de los resultados del reconocimiento y traducción del pīnyīn de los logogramas del lenguaje chino.

1.6 JUSTIFICACIÓN

La presente investigación contribuirá a un mejor aprendizaje del idioma chino. Y se aprovechará como proyecto base; para futuras investigaciones respecto a este idioma.

Esta investigación se desarrolla con el fin de mejorar el servicio ofrecido a la comunidad estudiantil, interesada en aprender una lengua extranjera como lo es el chino. Este trabajo apunta a contribuir a la calidad de educación a su mejora en oportunidades de trabajo y estudio.

La tecnología debe contribuir en mejora el aprendizaje de nuevas lenguas extranjeras.

1.6.1 JUSTIFICACIÓN SOCIAL

En el contexto nacional muchas personas se han dedicado a la importación y posterior comercio de productos chinos. Algunos de ellos, viajan incluso al país oriental teniendo que pasar las dificultades que implica el no conocer el idioma.
Un traductor básico les permitiría conocer la caducidad que tienen sus productos; que dicen los caracteres inscritos en los productos que importan.
Tener el conocimiento básico del idioma permitirá estrechar mucho más los lazos bilaterales entre los dos países.

1.6.2 JUSTIFICACIÓN TÉCNICA

Las técnicas de modelación, diagramas de transición y reconocimiento de cadenas de caracteres, independientemente de la lengua natural a la que pertenece; permiten tener una clara valoración del carácter (logograma) que es objeto de estudio.

1.6.3 JUSTIFICACIÓN ACADÉMICA

Para un estudiante egresado de la Universidad, hoy en día contar con conocimiento básico del idioma chino, le abre un sin fin de oportunidades académicas; para seguir ampliando su conocimiento en diferentes áreas de investigación científica.

1.7 LIMITES

El estudio realizado sobre la herramienta de apoyo a la traducción automática sólo será para usuarios interesados en la traducción de los caracteres (logogramas) chinos al español.

1.7.1 LIMITE TEMATICO

Este traductor solo traduce el significado de los logogramas de un curso de chino básico, como ser: los números, la hora y la fecha, los pronombres personales, parentesco familiar, los verbos y los tiempos, expresiones modales, palabras más útiles, etc. También frases y oraciones cortas. Ya que para interpretar la oración, se necesita un análisis sintáctico gramatical en el contexto, el cual por ahora no es objeto de nuestra investigación.

1.7.2 LIMITE ESPACIAL

La presente investigación es desarrollada en el contexto Latinoamericano más propiamente en Bolivia, ciudad de La Paz. Con la expectativa de contribuir al estudio básico del idioma chino.

1.7.3 LIMITE TEMPORAL

El periodo de la investigación durara de seis a diez meses; esto debido a que un curso básico de chino dura entre seis a siete meses. El análisis lexicográfico (lingüístico) e informático ira paralelo al curso después de transcurrido los primeros dos meses.

Cabe mencionar que el análisis informático, en este caso se hará desde el enfoque de un autómata formal traductor.

1.8 ALCANCES

La presente investigación contempla los siguientes alcances:

- Como alcance primordial tener un conocimiento básico del idioma.
- Se utilizara términos para un chino básico, específicamente para personas que estén aprendiendo el idioma chino.

- Para el desarrollo de los algoritmos de verificación de los logogramas se utiliza, la teoría de autómatas que nos permitirá verificar el verificar el pīnyīn asociado al logograma antes de traducirlo con las reglas de traducción.
- Cubrir las expectativas respecto a la aplicación de las técnicas que se utilizan para el desarrollo del prototipo.
- Comparación minuciosa de un carácter para su posterior traducción y así poder generalizarlo para los demás caracteres.

1.9 METODOLOGIA Y HERRAMIENTAS

1.9.1 METODOLOGIA DE LA INVESTIGACION

El proceso de investigación se realizara aplicando el método científico que involucra una serie de pasos:

Observación:

Consiste en la recopilación de hechos acerca de un problema ó fenómeno natural que despierta nuestra curiosidad. Las oraciones deben ser lo más claras y numerosas posible, porque han de servir como base de partida para solución.

Hipótesis:

Es la aplicación que nos damos ante el hecho observado. Su utilidad consiste en que nos proporciona una interpretación de los hechos de que disponemos, interpretación que debe ser puesta a prueba por observaciones y experimentos posteriores. Las hipótesis no deben ser tomadas nunca como verdaderas, debido a que un mismo hecho observado puede explicarse mediante numerosas hipótesis. El objeto de una buena hipótesis consiste solamente en

darnos una explicación para estimularnos a hacer más experimentos y observaciones.

Experimentación:

Consiste en la verificación o comprobación de la hipótesis. La experimentación determina la validez de las posibles explicaciones que nos hemos dado y decide el que una hipótesis se acepta o se deseche.

Teoría:

Es una hipótesis en la cual se han relacionado una gran cantidad de hechos acerca del mismo fenómeno que se nos da. Algunos autores consideran que la teoría no es otra cosa más que una hipótesis en la cual se consideran mayor número de hechos y en la cual la explicación que nos hemos forjado tiene mayor número de hechos y en la cual la explicación que nos hemos forjado tiene mayor probabilidad de ser comprobada positivamente.

Ley:

Consiste en un conjunto de hechos derivados de observaciones y experimentos debidamente reunidos, clasificados e interpretados que se consideran demostrados. En otras palabras la ley no es otra cosa que una hipótesis que ha sido demostrada mediante el experimento. La ley nos permite predecir el desarrollo y evolución de cualquier fenómeno natural.

1.9.2 HERRAMIENTAS

Para desarrollo del prototipo se utilizara el lenguaje de programación Visual. Net y SQL Server, sobre plataforma Windows.

Visual. Net

Visual. Net permite a los programadores crear robustas aplicaciones para Microsoft Windows y Windows NT incluye Microsoft Transaction Server, Internet Information Server además de los controles de cuadricula, de fichas y los controles enlazados a datos. Con las mejoras de productividad que brinda NET Framework de Microsoft y una experiencia de codificación muy productiva.

SQL Server 2008

SQL Server 2008 de Microsoft puede trabajar con un gran número de base de datos y cada una de ellas puede almacenar datos inter-relacionados ó datos no relacionados con los de las otras bases de datos.

Para lograr los altos niveles de rendimiento necesarios en sitios web muy grandes, un sistema de varios niveles permite equilibrar la carga de procesamiento de cada nivel entre varios servidores. SQL Server 2008 crea una partición horizontal de los datos de SQL Server para repartir la carga del procesamiento de la BD entre un grupo de servidores. Estos servidores se administran independientemente, pero cooperan en el procesamiento de las peticiones de BD de las aplicaciones.

SQL Server a demostrado su potencialidad y seguridad en los distintos sistemas grandes brindando una buen seguridad en los datos.

1.10 APORTES

1.10.1 APORTE TEÓRICO

Analizando esta propuesta de investigación sobre el desarrollo de un traductor de logogramas del idioma chino; se espera probar que la teoría de autómatas,

las aplicaciones OCR, son útiles a las necesidades de aplicaciones informáticas propuestas.

1.10.2 APORTE PRÁCTICO

Con el desarrollo del software de traducción automática de logogramas chinos se podrá contribuir a un mejor estudio del idioma; ya que entre más constante sea la práctica e inter-relación entre el usuario y el traductor, el aprendizaje será mayor.

2. MARCO TEÓRICO

2.1 LA LENGUA CHINA

A lo largo de la historia se han utilizado diversos estilos de grafismo que se conservan todavía en el arte caligráfico tradicional: El primer estilo de escritura, surgido durante la dinastía Shang, se conoce como dàzhuànshū 大篆書, "escritura del sello grande" otros estilos son la escritura del sello pequeño xiǎozhuànshū 小篆書, la escritura administrativa lishu 隸書, el xingshu 行書, y el de hierba cǎoshū 草書.

El estilo escrito o lengua literaria llamado *wenyan*, originado a partir del chino antiguo permite comprender los textos de cualquier época.

El estilo *baihua* más cercano al lenguaje hablado y utilizado ya por el budismo, desplazó progresivamente a *wenyan* y lo sustituyó completamente con la reforma educativa de China del siglo XX.

La forma tradicional de escribir era en vertical y de derecha a izquierda pero modernamente se hace en horizontal y de izquierda a derecha. A partir del 1956 el gobierno de la República Popular China simplificó los caracteres y adoptó oficialmente el año 1979 el sistema de transcripción pīnyīn.

La mayor parte del léxico chino moderno se compone de palabras bisílabas, entendiendo como palabra una unidad léxica que se puede combinar libremente en una frase. En el chino clásico se utilizaban muchas más palabras monosilábicas pero, aun así, no se sabe de ningún estado de la lengua en que todas las palabras hayan sido monosilábicas. De hecho, existen términos bisílabos que se escriben con dos caracteres que sólo pueden aparecer juntos, como por ejemplo *gāngà* (尷尬 /*尷尬*, "avergonzado") o *jǔyǔ* (齟齬 /*齟齬*, "altercado"). En estos casos, ni tan siquiera sería posible un análisis semántico o etimológico como unión de dos morfemas.

Cada signo ó logograma se refiere a la unidad mínima de significación (monema). En el léxico de la gran mayoría de las palabras actuales son bisílabos compuestos por unión de dos monemas, que habitualmente tienen identidad y significación propias, por ejemplo 歡迎 (bienvenida).

Pocos hanzi son ideogramas y desde tiempos antiguos los signos tienen valor fonético aunque estos signos no determinan exactamente el sonido y por lo tanto es muy difícil saber cómo deben ser pronunciados.

El número de caracteres chinos que figuran en el diccionario *Kangxi* es de 47.035, aunque muchos son variantes inusuales acumuladas a lo largo de la historia. Estudios llevados a cabo en China, han demostrado que la plena alfabetización en idioma chino sólo requiere conocer entre tres y cuatro mil caracteres.

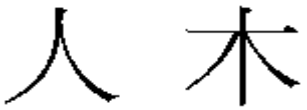
La mayoría de caracteres están compuestos por un radical semántico (de significado más o menos común) y uno fonético, lo que aprovecha para ordenarlos en los diccionarios.

Existen diversos criterios para clasificar los tipos de caracteres chinos. Lo más sencillo es dividirlos en tres categorías básicas:

1er TIPO

2.1.1 Caracteres pictográficos o pictogramas (象形 *xiànxíng*):

Los caracteres más antiguos son pictogramas, esto es, dibujos del concepto que representan. Por ejemplo:



El primer carácter, pronunciado *rén* en chino mandarín moderno, significa "persona", y procede del dibujo de un perfil humano. Este carácter es una auténtica palabra monosilábica y se utiliza en chino moderno. El segundo ejemplo, pronunciado *mù*, que significaba "árbol" en la antigüedad, y representa, de manera estilizada, el tronco, la copa y las raíces del árbol. En chino moderno, este carácter ha pasado a significar "madera", mientras que árbol se dice *shù* **(樹 /树).**

El pictograma son los que representan gráficamente a un objeto; son los más fáciles de reconocer. Ejemplos lo constituyen 口 *kǒu* “boca”, 日 *rì* “sol”. Dentro de este primer tipo esta también los caracteres indicativos o deictogramas (指事 *zhǐshì*): son símbolos que indican ciertos fenómenos y representaciones. Se emplean también para representar ideas abstractas.

A esta categoría pertenecen caracteres como 上 *shàng* “arriba” o 下 *xià* “abajo”. También lo es 本 *běn* “raíz” u “origen”, carácter que se obtiene añadiéndole un trazo a 木 *mù* “árbol” en su parte inferior, con lo cual se indica a sus raíces.

2do TIPO

2.1.2 Caracteres asociativos o ideogramas (會意 *huìyì*):

El segundo tipo de caracteres son los llamados ideogramas, son los creados por la asociación de dos o más caracteres pictográficos En estos casos los pictogramas se combinan para sugerir ideas por asociación. Por ejemplo:

Estos dos ideogramas se basan en los pictogramas anteriores. El primero, pronunciado *qiú*, significa "prisionero", significado sugerido por la imagen de una persona encerrada. En chino moderno, la palabra normal para decir prisionero es *qiúfàn* (囚犯), forma bisílaba que aún contiene este carácter.

El segundo carácter de la imagen significa "bosque", idea sugerida por la repetición del árbol. En este caso, el chino moderno también ha acabado dándonos una forma bisílaba: La palabra actual es ***sēnlín*** (森林), donde aparece también otro ideograma similar con tres árboles.

En este caso, cada carácter aporta un sema o unidad de sentido diferente, por lo que podrían denominarse caracteres de montaje sémico (lo cual seguiría el sentido de su nombre en chino, 會意 o “composición de significados”). Un ejemplo lo constituye el carácter 明 *míng* “brillante”, ensamblado a partir de 日 *rì* “sol” y 月 *yuè* “luna”, los dos cuerpos celestes más brillantes visibles en el firmamento. O 鳴 *míng* “voz o chillido de un ave”, compuesto de 口 *kǒu* “boca” y 鳥 *niǎo* “ave, pájaro”.

3er TIPO

2.1.3 Caracteres pictofonogramas o logogramas (形聲 *xíngshēng*):

El tercer tipo de caracteres lo constituyen los logogramas. Este tipo abarca la inmensa mayoría de los caracteres chinos actuales. Consiste en la modificación de otro carácter con el que comparte pronunciación añadiéndole otro componente que lo distingue. El componente añadido es a menudo uno de los llamados radicales, que aporta una idea semántica respecto al tipo de significado representado por el nuevo carácter. Veamos dos ejemplos:

Estos dos logogramas están basados en los ideogramas anteriores, pero corresponden a palabras totalmente diferentes. En ambos caracteres se aprecian tres trazos a la izquierda. Estos trazos son conocidos como "tres gotas de agua", o "radical del agua", y proceden del pictograma que significa agua. Los caracteres que tienen estas tres gotas de agua suelen tener un significado relacionado con el agua o los líquidos. El primero, pronunciado *qiú*, se basa en el ideograma *qiú* por el mero hecho de que tiene la misma pronunciación. Su significado clásico es "nadar" y se utiliza poco en chino moderno. Una palabra con este carácter es *qiúdù* (泅渡, "cruzar a nado"). El segundo carácter se pronuncia *lín*, y es por esa coincidencia fonética por la que se basa en el carácter del bosque. Las tres gotas de agua nos indican que se trata, sin embargo, de un término relacionado con el agua. Su significado es "empapar". En chino moderno se puede utilizar como verbo monosílabo, o en algunas combinaciones bisílabas, como en la palabra *línyù* (淋浴, "ducha"). Estos caracteres son los creados por la asociación de un carácter que aporta el elemento fonético y otro que añade la diferenciación semántica, llamado radical y que normalmente va hacia la izquierda. Este método de montaje

fonosemántico fue el que disparó las posibilidades combinatorias de la escritura china, resultando en que la mayor parte de los caracteres chinos actuales son de este tipo. El principio funciona del siguiente modo: si tengo el carácter 妈 *mā*, sé que algo perteneciente a la categoría 女 *nǚ* “mujer” se pronuncia como 马 *mǎ*, siendo aquí irrelevante el sentido de “caballo” de este último carácter. El significado de 妈 *mā* sería entonces “mamá”, empleándose también reduplicado, 妈妈 *māma*. La fórmula quedaría así: RADICAL (女 *mujer*) + FONÉTICO (马 *mǎ*) = SER FEMENINO QUE SE PRONUNCIA [ma] →妈 *mā* “mamá”.

Lo más probable es que en un estadio antiguo de la lengua estos logogramas empezaran escribiéndose con el mismo carácter cuyo sonido comparten, y que el añadido del radical se produjera posteriormente para clarificar el significado, de manera análoga, salvando las distancias, similar al uso que hacemos en castellano de la tilde para diferenciar monosílabos de significado diferente, como "si" y "sí", o "te" y "té". Por otro lado existen otras divisiones pequeñas como son:

Caracteres prestados (假借 *jiǎjiè*): se da sobre todo en palabras homófonas, en la que una presta a la otra el carácter, añadiéndose a veces otro elemento para distinguir los significados. Un ejemplo clásico es el verbo 來 *lái* “venir”, que se representa con el carácter que indicaba igualmente un tipo de cebada, lexía que igualmente se pronunciaba *lái*. El principio del préstamo de carácter para transcribir algo de difícil representación, por ser un concepto abstracto, como “venir”, o también por ser una palabra extranjera, lo encontramos en la transcripción de vocablos foráneos. Un antiguo ejemplo es 葡萄 *pútáo* “uva”.

Caracteres notativos: son aquellos en los que se ha ampliado el significado para abarcar otros conceptos semejantes. Cada carácter es indivisible e

invariable, pero por razones técnicas y de estudio se clasifican en simples y compuestos (según los trazos), ideográficos (sentido deducible de las partes) y fonético (si contienen alguna probable indicación sobre la pronunciación).

Lista de radicales: Los radicales entre paréntesis son versiones alternativas del radical anterior, total: 214 radicales.

- 1 trazo (6 radicales): 一 丨 丶 丿 乙 亅
- 2 trazos (23): 二 亠 人 儿 入 八 冂 冖 冫 几 凵 刀 力 勹 匕 匚 匸 十 卜 卩 厂 厶 又
- 3 trazos (31): 口 囗 土 士 夂 夊 夕 大 女 子 宀 寸 小 尢 尸 屮 山 川 工 己 巾 干 幺 广 廴 廾 弋 弓 彐 彡 彳
- 4 trazos (35): 心 戈 **戶** 手 支 攴 文 斗 斤 方 不 日 曰 月 木 无 止 歹 殳 毋 比 毛 氏 气 水 火 爪 父 爻 爿 片 牙 牛 犬
- 5 trazos (23): 玄 玉 瓜 瓦 甘 生 用 田 疋 疒 癶 白 皮 皿 目 矛 矢 石 示 禸 禾 穴 立
- 6 trazos (29): 竹 米 糸 缶 网 羊 羽 老 而 耒 耳 聿 肉 臣 自 至 臼 舌 舛 舟 艮 色 艸 虍 虫 血 行 衣 襾
- 7 trazos (20): 見 角 信 谷 豆 **豕** 豸 貝 赤 走 足 身 車 辛 辰 辵 (辶) 邑 酉 釆 里
- 8 trazos (9): 金 長 門 阜 隶 隹 雨 靑 非
- 9 trazos (11): 面 革 韋 韭 音 頁 風 飛 食 首 香
- 10 trazos (8): 馬 骨 高 髟 鬥 鬯 鬲 鬼
- 11 trazos (6): 魚 鳥 鹵 鹿 麥 麻
- 12 trazos (4): **黃** 黍 黑 黹
- 13 trazos (4): 黽 鼎 鼓 鼠
- 14 trazos (2): 鼻 齊

- 15 trazos (1): 齒
- 16 trazos (2): 龍 龜
- 17 trazos (1): 龠

Fuente: [Universidad de Granada(El chino de hoy)]

Continuando con nuestra explicación acerca de la lengua china es necesario e imprescindible hablar del pīnyīn.

2.2 EL PĪNYĪN

El **pīnyīn** es el sistema de transcripción oficial para adaptar las grafías chinas al alfabeto latino, a partir de la pronunciación del chino mandarín. Se han utilizado muchos sistemas de trascripción para aprender a pronunciar el chino. Actualmente la trascripción oficial adoptada a nivel internacional es el alfabeto pīnyīn, desarrollado en China a finales de los años 50.

Su función es que los antropónimos y **topónimos chinos** se escriban igual en todas las lenguas cuyo alfabeto sea el latino.

Sistema Pīnyīn

- Hay que tener en cuenta que las grafías *pinyin* tienen un sonido propio que no se corresponde con el sistema español, inglés o de otras lenguas.
- El chino es una lengua tonal y es difícil saber cómo acentuar cada palabra, por lo que la forma más fácil de resolver este problema es acentuando cada sílaba de la palabra.

1. La transcripción *pinyin* no admite la omisión de ningún carácter del nombre transcrito para simplificarlo, pues puede inducir a una confusión debido a la correspondencia con otra palabra.

- Una misma grafía del *pinyin* puede corresponderse con distintos sonidos según su posición en la sílaba en la que se encuentra.

2.2.1 LOS TONOS DEL PĪNYĪN

En chino mandarín existen cuatro tonos y uno neutro o tono ”ligero”. Cada sílaba pronunciada está afectada por un tono, por una melodía de tono. Esta es una característica **fundamental** de la lengua china hablada.

Los cuatro tonos:

Un guión **“-”** indica el primer tono (altura mínima de voz).
Un acento agudo “ ´ ” indica el segundo tono (altura media-baja de voz).
Un acento circunflejo invertido “ ˇ ” indica el tercer tono (altura media de la voz).
Un acento abierto “ ` ” indica el cuarto tono (altura alta de la voz).
Los acentos se sitúan sobre la vocal principal de cada sílaba.

Por ejemplo, la silaba **“ma”**. Esta se puede pronunciar:

1) En primer tono: **mā** ; se pronuncia (maaa) el carácter mā 妈 que significa mamá.
2) En segundo tono: **má** ; se pronuncia (maá) el carácter má)林 que significa “cáñamo”.
3) En tercer tono: **mă** ; se pronuncia (maaa) en la “a” del centro se baja el tono el carácter mă 马 que significa “caballo”.
4) En cuarto tono: **mà** ; se pronuncia (má!) el carácter mà que significa “injuriar”.
5) El tono neutro: **ma** ; se pronuncia (ma) es una (partícula interrogativa) el carácter ma 吗 que significa “¿qué…?, ¿Cómo?”, etc.

Puntualizamos, que no existe solo un carácter “mamá” que se pronuncie **ma** en primer tono ni solo el carácter “caballo” para el tercer tono.

Para el primer tono podemos encontrar, mā (un tipo de rana) que se escribe con distinto carácter chino, entre otros.

Es esencial acostumbrarse a hacer distinción entre los cuatro tonos. Ello implica repetidos ejercicios de pronunciación.

Aprendiendo los cuatro tonos, se recuerda que el tercer tono es siempre más acentuado que los otros, pronunciación aproximada (maa). También, las vocales simples son más acentuadas cuando están solas en una sílaba y mas atenuadas cuando forman parte de un grupo de vocales en una misma sílaba [Assimil, 2011].

2.3 RECONOCIMIENTO OPTICO DE CARACTERES (OCR)

La tecnología de reconocimiento de caracteres, OCR **(Optical Character Recognition)** engloba un conjunto de técnicas basadas en física, matemática y estadística, en donde las formas de los caracteres, sus transformaciones y comparaciones, que complementándose entre sí, se emplean para distinguir de forma automática entre los diferentes caracteres alfanuméricos.

El OCR puede leer virtualmente cada clase de texto impreso, ejemplo: dígitos, símbolos, letras mayúsculas, letras minúsculas, letras acentuadas, puntuación. Las facultades más grandes del OCR es su capacidad de incorporar datos en una base de datos y su exactitud de lectura demostrada. [Green P, 2000]

El OCR e ICR son herramientas muy usadas en informática; estos son métodos automatizados de la recogida de datos ampliamente utilizados, para procesos de alta velocidad.

Se puede proporcionar métodos de proceso del documento OCR de varios formatos, como ser:

- Reconocimiento de cheques con el OCR.
- Reconocimiento de texto completo con el OCR, etc.
- Reconocimiento de texto manuscrito con ICR, etc.

VISIÓN GLOBAL DE OCR

En todo sistema de reconocimiento óptico de caracteres (OCR) se distinguen al menos estas 4 etapas:

- Adecuación de la imagen **(preproceso).**
- Selección de la zona de interés **(segmentación).**
- Representación digital de la imagen **(extracción de características).**
- Distinción del carácter contenido en la imagen **(reconocimiento).**

Para cada una de las cuatro etapas es posible aplicar multitud de técnicas ya existentes o desarrollar alguna específica en función de las condiciones en las que se presentan los datos de entrada, que en el caso de OCR se puede traducir por las imágenes de entrada.

2.3.1 PREPROCESAMIENTO

En esta fase el objetivo que se persigue es eliminar de la imagen de cualquier tipo de ruido o imperfección que no pertenezca al carácter, así como normalizar el tamaño del mismo. Además, para el caso de OCR, la normalización de la imagen también puede implicar un binarizado de la misma.

Para la eliminación del ruido que puede aparecer en una imagen digital, bien provocado por manchas reales o grafías imperfectas, o bien por defectos técnicos en la adquisición o binarizado de la imagen, se utilizan diversos algoritmos:

- **Etiquetado:** para la división de la imagen en regiones de componentes conectadas.
- **Erosión / expansión:** para la eliminación de pequeños grupos de píxeles.
- **Umbralizado de histograma:** para eliminar/seleccionar los objetos más brillantes o más oscuros de la imagen.

2.3.2 SEGMENTACIÓN

Una vez preprocesada la imagen se deberá fragmentar o segmentar en las diferentes componentes conexas (parte de la imagen donde todos los píxeles son adyacentes entre sí) que la componen. La fragmentación o segmentación de la imagen constituye una de las mayores dificultades del reconocimiento, y se hace necesaria para poder reconocer cada uno de los caracteres de la imagen binaria.

Para reconocer caracteres es necesaria, en primer lugar, su localización dentro del texto del documento, teniendo en cuenta, en esta operación el orden en el que se disponen en el mismo y los espacios en blanco y finales de línea, para que pueda recomponerse el texto tal y como se encontraba en el documento original.

Existen tres magnitudes que determinan el orden de los caracteres dentro de un texto: los renglones de los que consta, las palabras de un renglón y las letras de una palabra.

A la hora de segmentar un texto lo primero que se hace es detectar los distintos renglones que forman el texto. Para conseguirlo se realiza el siguiente procedimiento:

1. Se hace una proyección horizontal (histograma) consistente en contar los elementos de tinta que existentes en cada una de las filas, traspasando estos valores a otra matriz, unidimensional, resultado de la proyección, en la que existirán diferentes zonas de densidad de tinta separadas por otras vacías. Cada zona donde la proyección dé un valor no nulo será interpretado como un hipotético renglón.

2. Se analiza la matriz unidimensional para detectar los posibles renglones de los que está compuesto el texto. Si se detecta una línea con densidad de proyección no nula y además la anterior estaba en blanco, en esa línea comienza un renglón. A continuación se realiza la misma operación pero a la inversa, se busca la línea posterior que sea blanca y que la anterior no lo fuera, ahí estará el final del renglón. Este método se aplica sucesivamente hasta el final de la matriz de proyección, consiguiendo así delimitar los renglones que forman el texto.

2.3.3 EXTRACCIÓN DE CARACTERÍSTICAS

Desde el punto de vista del reconocimiento de formas, la matriz bidimensional se ve como un vector de tantas dimensiones como componentes tiene la matriz. La dimensión de estos vectores (el número de componentes) es normalmente elevado, lo que supone un gran coste computacional a la hora de procesar el mismo. Y no solo eso, y más importante aún, es que está comprobado que al intentar clasificar ("reconocer") vectores de este tamaño aparece un efecto, llamado maldición de la dimensionalidad, que provoca que los resultados, independientemente del método de clasificación utilizado, no sean satisfactorios. Por ello se han desarrollado multitud de técnicas, denominadas "técnicas de selección y extracción de características", mediante las cuales es posible obtener una representación del objeto a reconocer más eficiente.

Eficiencia, en este caso, significa que con una representación más compacta se consigue un poder discriminativo igual o superior al que se tenía con la representación original. Esto no es solo importante por el ahorro de espacio en el almacenamiento de las muestras, sino que durante el proceso de reconocimiento reduce los costes computacionales, debido a la reducción en el volumen de información procesado.

La extracción de las características es una de las fases más difíciles en los sistemas de reconocimiento de caracteres, puesto que es muy difícil escoger un conjunto de características óptimo.

Para que una característica se pueda considerar buena debe poseer:

- Discriminación: Deben ser características que diferencien suficientemente una clase de otra.
- Deben tener igual valor para mismas clases.
- Independencia: Las características deben estar no correlativas unas de otras.
- Pequeño espacio para características: El número de características debe ser pequeño para la rapidez y facilidad de clasificación.

En el campo de investigación del reconocimiento de formas se tiene experiencia en el uso de algunos métodos de extracción de características basados en transformaciones del espacio de representación de las muestras. Algunos de estos son: PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), ICA (Independent Component Analysis), NDA (Non-linear Discriminant Analysis).

2.3.4 RECONOCIMIENTO

Una vez que se tienen las características más importantes de la imagen a analizar, hay que determinar el carácter correspondiente. Existen diferentes enfoques a la hora de generar modelos matematicos para el reconocimiento de patrones. Como ser:

KNN

Para OCR, este método es muy conveniente, no paramétrico y supervisado, que proporciona resultados muy adecuados para la aplicación que se está tratando, El algoritmo K-NN (K vecinos más próximos). Este método es muy

popular debido a su sencillez y a cierto número de propiedades estadísticas bien conocidas que le proporcionan un buen comportamiento para afrontar diversos tipos de problemas de clasificación, siendo uno de ellos el de OCR.

Árboles de decisión

Los árboles de decisión, es una técnica de minería de datos que se puede aplicar en el contexto de reconocimiento óptico de caracteres. Su aprendizaje es inductivo y no supervisado. Los patrones o atributos que se quieren evaluar de un carácter determinado constituyen los nodos del árbol, mientras que los resultados finales de los mismos se almacenarán en las hojas del mismo. Tras la construcción del árbol y dada la estructura del mismo, toda la evaluación de caracteres se puede tratar como una arquitectura IF-THEN ELSE, por lo que si el número de parámetros a evaluar es suficientemente grande para tener capacidad expresiva pero suficientemente pequeño para ser eficiente computacionalmente, el árbol resulta una estructura que favorece mucho la velocidad de cálculo, no como el algoritmo K-NN que aunque es muy conveniente, puede llegar a ser muy costoso computacionalmente.

Redes neuronales

Las redes neuronales son esquemas de minería de datos que intentan imitar la arquitectura del cerebro. Se componen de una serie de unidades básicas, llamadas neuronas, que básicamente reciben una entrada, la multiplican por unos pesos y presentan una salida con una función de ajuste, que depende de la suma de salidas de la etapa anterior.

Sirven para representar y ajustar muy eficazmente cualquier función que sería muy difícil de definir en términos algebraicos. Un esquema funcional de una red de neuronas podría ser el siguiente:

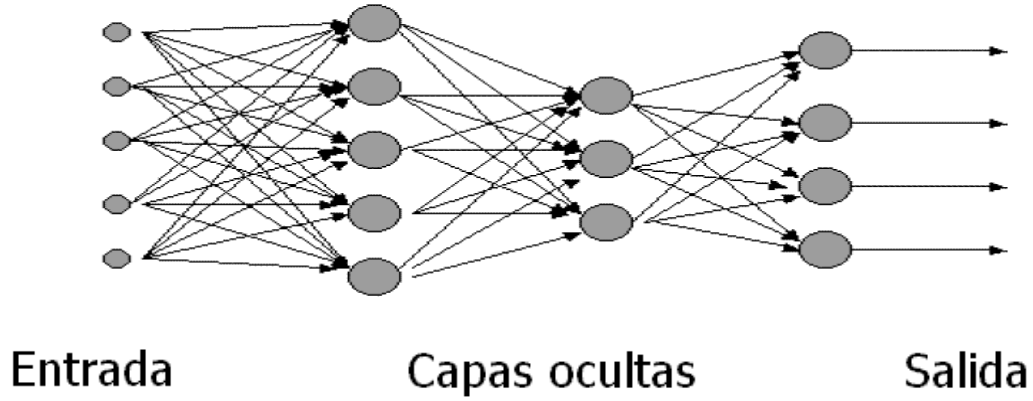


Figura 2.1

Fuente [Universidad Carlos III, Victor Sandonis]

La topología de las redes puede ser muy variada y según esta topología las redes se pueden clasificar en:

- Feed Fordward: Red clásica, las salidas alimentan las entradas de la etapa siguiente, en la red no se permite la aparición de ciclos o realimentaciones.
- Feed Back: En esta versión si se permiten ciclos cerrados dentro de la red
- Lateral: Adicionalmente, se permite la comunicación vertical entre neuronas de la misma capa, además de en la dirección horizontal, como en las redes clásicas.

2.4 RECONOCIMIENTO DE CARACTERES INTELIGENTES (ICR)

Esta tecnología avanzada, de la exploración puede traducir una variedad amplia de las fuentes y del tipo de estilos impresos, de las fuentes de papel al texto electrónico, permite a sistemas de exploración y de proyección de imagen dar vuelta a imágenes de caracteres impresos a mano en caracteres legibles por la maquina.

Los sistemas ICR llevan a los OCR un paso adelante utilizando programas de computo para aplicar pruebas de inteligencia lógica a los caracteres escaneados para convertirlos de manera más confiable en información más legible para la computadora.

Los sistemas de ICR aplican reglas de ortografía, gramática y contexto para escanear los textos a fin de efectuar evaluaciones “inteligentes” sobre la interpretación correcta de la información. Esto permite una conversión mucho más precisa de los textos escaneados, de la que realiza los sistemas OCR más simples, especialmente con el texto manuscrito.

Los programas ICR requieren computadoras rápidas y poderosas para desempeñar de manera eficiente. Los sistemas ICR de alta confiabilidad solo estuvieron disponibles a mitad de la década de 1990 con el desarrollo de productos computacionales, económicos y poderosos.

A medida que se vuelvan mas confiables los sistemas ICR, se incrementaran sus aplicaciones electorales. Son particularmente apropiados para capturar información de formatos. También se está evaluando su capacidad para capturar números manuscritos de las papeletas que utilizan sistemas electorales más complejos, como el voto alternativo ó el de voto único transferible. A la fecha, los sistemas automatizados de captura de información no han sido utilizados para estos sistemas electorales debido a la complejidad de la tares [Green p., 2000]

2.5 DEFINICIÓN DE AUTÓMATA FINITO

Un autómata finito o máquina de estado finito es un modelo matemático de un sistema que recibe una cadena constituida por símbolos de un alfabeto y determina si esa cadena pertenece al lenguaje que el autómata reconoce.

Éstos se definen mediante una quíntupla $(\sum, Q, f, q_o, F)$ donde:

$\sum$: alfabeto de entrada.

Q: conjunto de estados; es conjunto finito no vacío.

f: función de transición. f(p,a)=q

q_o : (perteneciente a Q) estado inicial.

F : (perteneciente a Q) conjunto de estados finales o de aceptación.

2.5.1 CLASIFICACIÓN

Los autómatas se pueden clasificar en:

Deterministas

- Cada combinación (estado, símbolo de entrada) produce un solo estado.

No Deterministas

- Cada combinación (estado, símbolo de entrada) produce varios estados y además son posibles las transiciones con λ.

2.5.2 REPRESENTACIÓN

Los autómatas se pueden representar mediante tablas de transición o diagramas de transición.

Tablas de transición:

• Filas encabezadas por los estados (Q)

• Columnas encabezadas por los símbolos de entrada (∑)

TABLA	a	b
→ p	q	
*q	q^3	r
q^3		r
*r	q^3	

Diagramas de transición:

- Nodos etiquetados por los estados(Q)
- Arcos entre nodos etiquetados con (∑)
- q_o se señala con →
- El estado final se señala con * o con doble circulo.

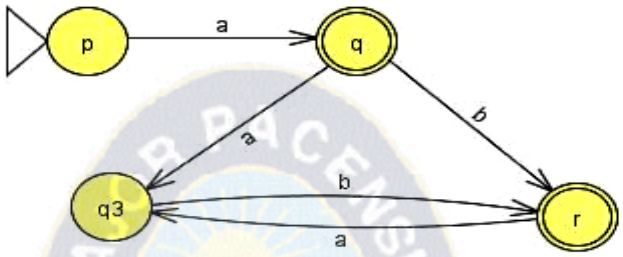


Figura 2.2

Fuente[Yannis Haralambous "Fonts & Encodings", 2007]

2.5.3 DESCRIPCIÓN DE UN AUTÓMATA FINITO DETERMINISTA

Consideremos el lenguaje regular ∑ representado por *c*(a U bc*)*.* Si dada una cadena ***w*** se nos pregunta si ***w*** pertenece a ∑, debemos analizar no solo los caracteres que aparecen en ***w,*** sino también sus posiciones relativas. Por ejemplo, la cadena $abc^5\ c^3\ ab$ está en ∑, sin embargo $cabac^3\ bc$ no lo está. Podemos construir un diagrama que nos ayude a determinar los distintos miembros del lenguaje.

Tal diagrama tiene la forma de un grafo dirigido con información adicional añadida, y se llama ***diagrama de transición***. Los nodos del grafo se llaman ***estados*** y se usan para señalar, en ese momento, hasta que lugar se ha analizado la cadena.

Las aristas del grafo se etiquetan con caracteres del alfabeto y se llaman ***transiciones***. Si el siguiente carácter a reconocer concuerda con la etiqueta de

alguna transición que parta del estado actual, nos desplazamos al estado que nos lleve a la arista correspondiente. Naturalmente, nosotros debemos comenzar por un estado inicial, y cuando se haya tratado todos los caracteres de la cadena correspondiente, necesitamos saber si la cadena es “legal”.

Para ello se marcan ciertos estados como ***estados de aceptación ó finales***.
Sí cuando ha sido tratada la cadena en su totalidad terminamos en un estado de aceptación, entonces la cadena es “legal “. Marcaremos el estado inicial con una flecha ( ) y alrededor de los estados de aceptación trazaremos un círculo. [Dean Kelley; 1995].

Por ejemplo: Sea el AFD1 = ({a,b}, {p,q,r}, f, p, {q}) donde f está definida por:
f(p,a) = q f(p,b) = r
f(q,a) = q f(q,b) = r
f(r,a) = r f(r,b) = r

Escribir su tabla de transición y dibujar su diagrama de transición.
estados (Q): p, q, r
estado inicial: p
estado final: q
símbolos: a,b

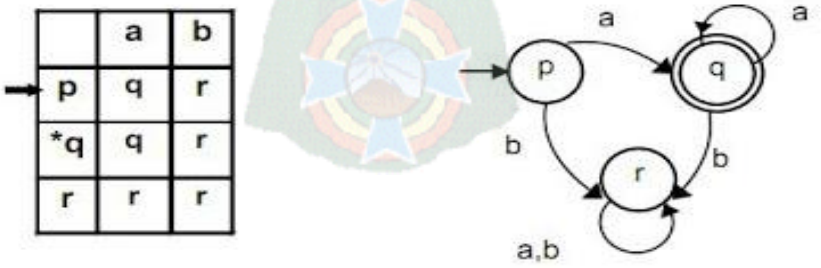


Figura 2.3
Fuente[Yannis Haralambous “Fonts & Encodings”, 2007]

2.5.4 DESCRIPCIÓN DE AUTÓMATA FINITO NO DETERMINISTA

Si tratamos de definir el término autómata finito no determinista, veremos que la mayor parte de la definición se puede obtener a partir de la de AFD. Es decir, tendremos un conjunto finito de estados Q, un alfabeto de entrada ∑, un estado inicial o de partida q_o, un conjunto de estados de aceptación F y una regla de transición. La única diferencia que existe se encuentra en las reglas de transición. En un AFN, las reglas asocian pares (q, σ) con colecciones o conjunto de estados. Esto significa que la regla es una relación entre Q x ∑ y Q, o sobre (Q x ∑) x Q. por tanto, definiremos un autómata finito no determinista mediante una colección de cinco objetos (Q, ∑, q_o, F, f), donde:

Q es un conjunto finito de estados.
∑ es el alfabeto de entrada.
q_o es uno de los estados de Q designado como estado de partida
F es una colección de estados de aceptación o finales.
f es una relación sobre (Q x E) x Q y se llama relación de transición.

	a	b	λ
→ *p	q		
q		p,r	s
r		p,s	r,s
s			r

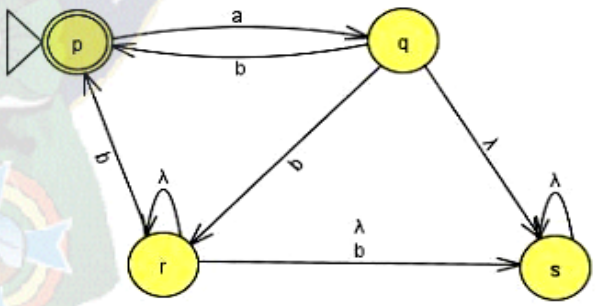


Figura 2.4
Fuente[Yannis Haralambous "Fonts & Encodings", 2007]

2.5.5 AUTOMATA FINITO COMO RECONOCEDOR DE LENGUAJE:

Cuando un AF transita desde q_O a un estado final en varios movimientos, se ha producido el RECONOCIMIENTO o ACEPTACIÓN de la cadena de entrada.

Cuando un AF no es capaz de alcanzar un estado final, se dice que el AF NO RECONOCE la cadena de entrada y que ésta NO PERTENECE al lenguaje reconocido por el AF, ejemplo:
A partir del autómata, comprobar si reconoce la palabra “aabbaba”.

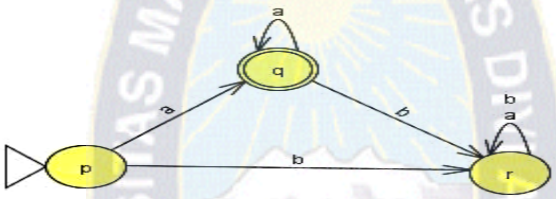


Figura 2.5
Fuente[Yannis Haralambous “Fonts & Encodings”, 2007]

Empezaríamos en el estado P, al recibir una "a", pasaríamos a Q, al recibir una "a" de nuevo, seguiríamos en Q, después al recibir una "b", pasaríamos a R, y recibiendo otra "b", seguiríamos en R. A continuación recibiríamos una "a", que nos llevaría a R y así hasta el final. Por lo tanto al quedar después de la palabra leída en el estado R, y dado que este no es un estado final como lo es Q, la palabra no se reconocería. En cambio la palabra “aa” si que se reconocería puesto que terminaríamos en el estado final Q. Se trata por tanto de un autómata que reconoce lenguajes formados exclusivamente por una o varias “a”.

2.5.6 AUTÓMATAS CONEXOS

Sea un AFD = ($\sum$, Q, f, q_O, F), el estado p es ACCESIBLE desde q si f'(q, x) = p. En otro caso se dice que es INACCESIBLE.

Resultado: Todo estado es accesible desde sí mismo pues f'(p, λ) = p

Por tanto, un autómata es conexo si todos los estados son accesibles desde el estado inicial. Para convertir un autómata no conexo en conexo hay que eliminar los estados no accesibles. Es importante destacar que un autómata conexo y su no conexo equivalente, reconocen el mismo lenguaje.

Conexo

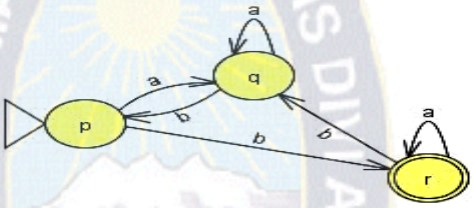


Figura 2.6

No conexo

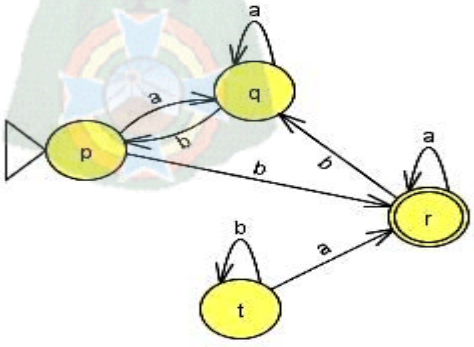


Figura 2.7

Fuente[Yannis Haralambous "Fonts & Encodings", 2007]

MINIMIZACIÓN

Si queremos calcular el autómata mínimo equivalente de cierto autómata, tenemos que obtener el conjunto cociente Q/E.

Para ello, asignamos clases a los diferentes estados, siendo para la primera iteración, C1(clase 1) para los estados finales y C2(clase 2) para el resto de estados.

Q/E_o = { C1(p, q, r) , C2(s, t, u, v)}

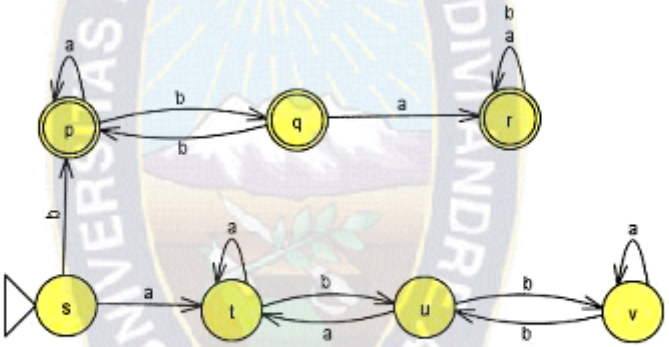


Figura 2.8

Fuente[Yannis Haralambous “Fonts & Encodings”, 2007]

2.5.7 AUTÓMATA FINITO TRADUCTOR

Un autómata finito traductor MT es simplemente un autómata finito que se define como una 7-upla:

MT = <Q, ∑, δ, q_0, F, S, t>.

Donde:

Q: Conjunto finito de *estados*,

∑: Alfabeto o conjunto finito de símbolos de entrada,

δ: Es la *función de transición de estados* definida **δ: Q x ∑ →*Q***

q_0: *Estado inicial* **q_o ∈ ∑.**

F: Conjunto de *estados finales o estados de aceptación.* **F⊆Q**.

***S:** Alfabeto* o conjunto finito de símbolos de salida.

t: Es la *función de traducción* definida **t: Q x ∑ → *S****

Ambas funciones **δ:** Q **x** ∑ → *Q* y **t:** Q **x** ∑ →*S** están definidas sobre Q **x** ∑.

Si existen **δ(**qi , *a***) =** *q*j y **t (**qi , *a***) =** *x*

donde qi, qj ∈ Q; *a* ∈ ∑; *x* ∈ S*

en el diagrama de transición de estados el valor de la traducción *x* se agrega sobre los arcos.

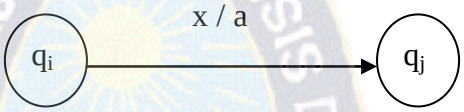


Figura 2.9

Fuente[Elaboración propia]

2.5.7.1 FUNCIÓN DE TRADUCCIÓN PARA CADENAS

La extensión de la función de traducción **T*: Q x ∑* → S***, tal que **T*(qi ,ω)** es la cadena que traducirá el autómata, luego de leer la cadena **ω** en la cinta de entrada y comenzando en el estado qi, se define como:

1) t*(qi, λ) = λ

2) t*(qi, ω*a*) = t*(qi, ω). t(δ*(q_i, ω),*a*) donde *a* ∈ ∑**,** ω ∈ ∑***,** q_0 ,qi ∈ Q.

La diferencia entre **t** y **t*** es que **t** se define desde un estado y un símbolo del alfabeto, y **t*** se define desde un estado y una cadena de símbolos.

2.5.7.2 TRADUCCIÓN

El autómata solo define la traducción, si el autómata finito reconocedor subyacente "acepta" la cadena. Es decir, la traducción **T(ω):** $\sum^* \rightarrow S^*$ asociada a M_T está definida como: **T(ω)=** $t^*(q_0 , \omega) \Leftrightarrow \delta^*(q_0 ,\omega) \in F$ donde $\omega \in \sum^*$

2.6 MODELO DE LENGUAJE

Un modelo de lenguaje es un mecanismo para definir la estructura del lenguaje, es decir, para restringir adecuadamente las secuencias de unidades lingüísticas más probables.

En general son útiles en aplicaciones que exhibían una sintaxis y/o semántica completa, un buen modelo de lenguaje solamente debería aceptar frases correctas y rechazar aquellas secuencias de palabras incorrectas.

2.7 TIPOGRAFIA

La tipografía trata de las letras, números y símbolos de un texto impreso (ya sea sobre un medio físico ó electromagnético), tales como un diseño, su forma, su tamaño y las relaciones virtuales que se establecen entre ellos.

Macro tipografía: la macro tipografía se centra en el tipo de letra, el estilo de la letra y en el cuerpo de la letra.

Micro tipografía ó tipografía del detalle: El término Mikrotypografie (microtipografía) se aplico por primera vez en un discurso dado en la Sociedad Tipográfica de Munich. Se ha generalizado desde entonces en la literatura especializada. No obstante, se puede substituir así mismo por una palabra inglesa. Detailtypografie (tipografía del detalle). Comprende los siguientes rubros: la letra, el espacio entre letras, la palabra, el espacio entre las palabras,

el interlineado y la columna, además de otras importantes funciones el peso visual, el interletrado y el interlineado.

Tipografía de edición: Reúne las cuestiones tipográficas relacionadas con los grupos de caracteres, el tamaño de las letras, los espacios entre las letras y las palabras; el inter tipo e interlinea y la medida de línea y columna o caja, es decir aquellas unidades que conceden un carácter normativo.

Tipografía creativa: Esta contempla la comunicación como una metáfora visual, donde el texto no sólo tiene una funcionalidad lingüística, y donde a veces, se representa de forma gráfica, como si se tratara de una imagen.

2.8 LEXICOLOGIA

La palabra "lexicología" se compone de dos términos griegos: "lexis", que significa palabra, y "logos", que significa estudio. La Lexicología es la rama de la Lingüística que estudia las palabras, su clasificación y representación según alguna relación sistemática. Con la aplicación de la lexicología y lexicografía se ha podido crear los diccionarios con los que contamos hoy en día.

Por otro lado, la lexicología pretende hacer generalizaciones sobre la estructura del léxico, las relaciones entre palabras y el modo en como la lengua representa ciertas áreas semánticas. La lexicología trata asuntos como:

- El origen de las palabras (etimología), algo para lo que se requiere el auxilio de la lingüística histórica.
- Las relaciones entre conceptos y palabras (onomasiología y semasiología).
- La estructura de relaciones semánticas que se establecen entre las palabras que constituyen el léxico de una lengua.[Wikipedia,2000]

3. MARCO PRÁCTICO

3.1 INTRODUCCIÓN

Para cumplir el propósito de esta investigación, consideramos la entrada de datos de la forma off-line, introducción de datos mediante scanner como una imagen cualesquiera, para luego hacer un estudio de la misma cumpliendo los objetivos específicos planteados en el capítulo uno, de reconocimiento y posterior traducción de caracteres chinos, se describe la obtención de datos desde la perspectiva, off-line. Las dos formas de obtención de datos de entrada a un ordenador son:

- On-line: Los datos se obtienen en tiempo real mientras se escribe. En el reconocimiento “on-line” el escritor está conectado directamente por medio de un bolígrafo electrónico, lápiz óptico o dispositivos similares a un computador, esta escritura es registrada en función del tiempo real.
- Off-line: los datos son obtenidos por medio de escáneres, cámaras, etc. todas estas en forma de imagen. Ver (fig.3.1).

Tras analizar en el capitulo dos los métodos y técnicas que se utilizaran, tanto para el reconocimiento óptico de caracteres (OCR) y el diseño de un Automata. En este capitulo, se describira de manera más detallada el proceso OCR a utilizar haciendo referencia a la etapa **reconocimiento** basada en **redes neuronales** para la distinción del carácter contenido en la imagen.

Figura 3.1

Asi tambien esbosaremos el diseño del modelo del Automata finito traductor, que nos permitira validar el lenguaje formal a traducir.

3.2 ETAPAS BASICAS DEL SISTEMA (OCR) Se distinguen al menos 4 etapas

3.2.1 PREPROCESO

Lo que caracteriza al preproceso off-line, es que se obtiene el texto de manera indirecta, primero se escribe en papel o cualquier soporte físico, para luego ser digitalizado utilizando un escáner o una cámara. Una imagen bidimensional, f(x, y), es una función, $f : R \times R \rightarrow R$, que para cada punto del espacio bidimensional devuelve el nivel de luz de dicho punto.

Una imagen off-line es una función, I(x, y), donde el espacio bidimensional y los niveles de gris se han discretizado; I: {1, ・ ・,F}×{1, ・ ・ ,C}→{1, ・ ・ ,L}, donde L suele ser 255, y F,C son las filas y columnas de la matriz. A cada una de las celdas de esta matriz se la denomina píxel (es la forma abreviada de las palabras pixture elemen). El proceso de adquisición consiste pues en una doble discretización: del espacio y de los niveles de gris, o dicho de otro modo, es el proceso que permite transformar una imagen bidimensional en una imagen off-line.

Etapa 1: Entrenamiento

El primer paso es determinar el conjunto de caracteres que van a ser utilizados, y diseñar una imagen con texto que contenga un conjunto de ejemplos. Los puntos más importantes que se deben tomar en cuenta a la hora de crear una imagen de entrenamiento son:

- Se debe tomar como minimo 5 muestras para los caracteres especiales.
- Para los caracteres más frecuentes se deben crear, por lo menos de 2 a 3 muestras, dependiendo de la legibilidad.
- Se debe hacer frases que tengan significado, no se debe hacer de la siguiente manera:174”#$%@ ya que no da muchas posibilidades de conseguir buenas mediciones en los caracteres especiales.

Umbralización

La umbralización, o *thresholding* consiste en transformar una imagen digital en escala de grises en una imagen digital binaria. Los métodos de umbralización clasifican los puntos de la imagen en dos clases: los que pertenecen al fondo de la imagen o segundo plano, *background pixels* y los que pertenecen al texto o primer plano, *foreground pixels*. Esta clasificación está basada en la elección de un umbral que divide los puntos en dos clases, aquellos cuyo valor está por encima del umbral que se clasificarán como *background pixels* (píxeles con valores elevados codifican tonos más claros) y los que están por debajo como *foreground pixels.*

En binario se suele codificar los píxeles *foreground* (negros) como 1 mientras que los de *background* como 0.

Nivel de Ruido

Todos los métodos expuestos en el preproceso actuarán sobre la página completa. Así como las técnicas para intentar reducir el ruido que pueda llevar

consigo la imagen. Como ruido se entiende toda aquella (señal originaria que no aporta ninguna información), añadida a la señal que codifica la información.

Normalización a nivel de texto

Corrección del *slope*

El *slope* es la pendiente o inclinación que presenta la linea base sobre la que está escrita una palabra, o una secuencia de ellas, con respecto al eje de ordenadas. Una vez determinado el ángulo, la corrección consiste en aplicar una rotación de la imagen con el mismo ángulo que el de *slope* en sentido contrario. El primer paso para determinar el ángulo de *slope* es dividir cada frase en segmentos de frase. Este proceso no pretende segmentar el texto en palabras, sino dividirlo en aquellas secuencias de texto que estén muy cercanas entre sí.

Tambien siguen las tecnicas de nivel slant y la normalizacion de tamaño pero estas tecnicas son utilizadas para la inclinación y tamaño de texto manuscrito.

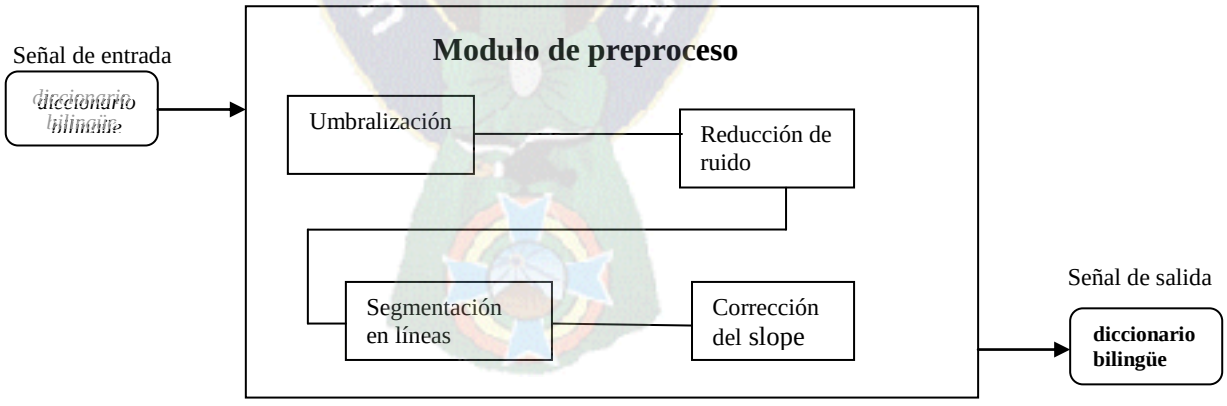


Figura 3.2 Esquema de preproceso con un ejemplo

Fuente [Elaboración propia]

3.2.2 SEGMENTACION

Una vez conocida la situación de los renglones y sus límites, se procede a aislar los caracteres. Como asumíamos inicialmente que no existiría solapamiento, se puede realizar una proyección vertical dentro de cada renglón para detectar los posibles caracteres. Esta proyección vertical (suma de los píxeles de cada línea con 1 para el píxel negro y 0 para el blanco) da un resultado nulo en las zonas donde no existe tinta, lo que representa la separación entre dos caracteres, y resultados no nulos que indican la presencia de caracteres. Los límites en altura, superior e inferior, se detectan analizando la vertical de cada carácter, desde la parte superior e inferior, respectivamente, del renglón que lo contiene. De esta forma, se consigue aislar cada carácter en una ventana rectangular con las dimensiones correspondientes, su anchura y altura. Este método, además, es válido aún en el caso de existir caracteres de distinto tamaño dentro del mismo texto. En cuanto a los caracteres "blancos" (espacios entre palabras), éstos se detectan cuando la separación entre dos caracteres consecutivos es mayor que un umbral dependiente de la altura del primero.

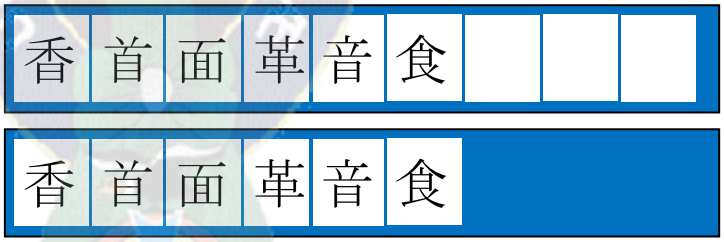


Figura 3.3

Fuente [Elaboración propia]

3.2.3 EXTRACCIÓN DE CARACTERÍSTICAS

Una vez realizada la segmentación, se tiene una imagen normalizada en la que se encuentra la información susceptible de ser “reconocida”. La extracción de características establece un nuevo espacio de representación de la señal de entrada. Este nuevo espacio debe permitir una representación compacta de la señal, y debe facilitar la discriminación entre las distintas clases de patrones, en el sentido de que se minimice la variabilidad intra-clase, al mismo tiempo que

maximiza la variabilidad inter-clase. Esto implica que los valores que tomen las características para las muestras de una clase presenten la menor variabilidad posible, al mismo tiempo que son distantes para el resto de muestras de las otras clases. Además, las características deben ser suficientemente invariables para que estén presentes en cualquier estilo de escritura.

La elección del conjunto de características es una decisión crítica que depende de la tarea y del clasificador que se use. El mejor conjunto de características para una tarea dada será aquel, entre aquellos adecuados para el tipo de reconocedor usado, que proporcione la máxima precisión con el mínimo número de características.

Además, es deseable que la extracción de características se obtenga de manera sencilla y a un bajo coste en recursos. Por lo tanto, un buen diseño debe maximizar la precisión del sistema y minimizar su tiempo de respuesta.

En esta tesis se ha utilizado un reconocedor de extracción de características basado en transformaciones del espacio de representación de las muestras.

LDA (Linear Discriminant Analysis):

Utilizan combinaciones lineales de variables para representar a los datos. En concreto, LDA modela **la diferencia entre las clases de datos.**

3.2.4 RECONOCIMIENTO

Una vez que se tienen las características más importantes de la imagen a analizar, hay que determinar el carácter correspondiente. Existen diferentes enfoques a la hora de generar modelos matematicos para el reconocimiento de patrones. Para esta investifacion se utilizara redes neuronales.

En el caso que nos ocupa, las redes de neuronas son una muy buena alternativa para la etapa de reconocimiento de caracteres, ya que una vez que se entrenan, y si este entrenamiento ha sido adecuado, pueden usarse para reconocer los caracteres que recibe como imágenes.

La versatilidad de las redes de neuronas permite extender su clasificación también a la naturaleza de sus datos y relacionados con estos, el tipo de aprendizaje que realizan. Por tanto, las redes que atienden a esta clasificación son:

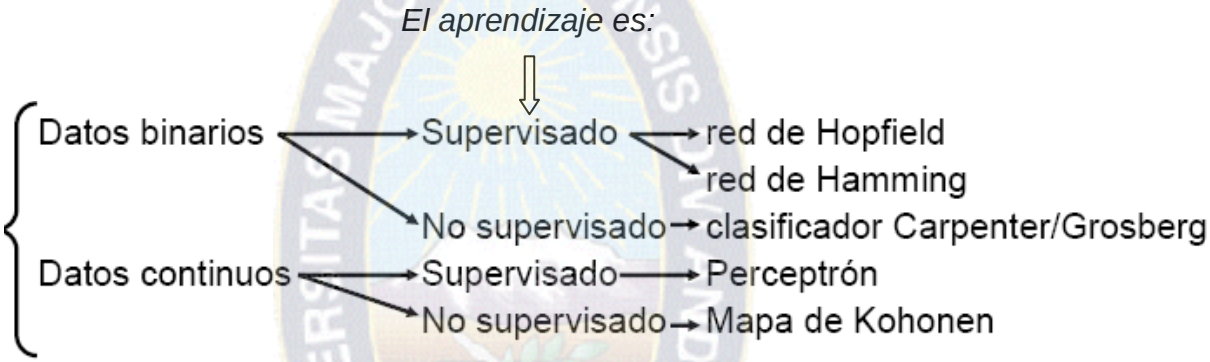


Figura 3.4 Algoritmo de proceso de reconocimiento

Fuente: [Wikipedia]

Por lo tanto, supondremos que partimos de una imagen de un texto escrito, organizado en renglones. La estrategia de resolución del problema de OCR tendrá los siguientes pasos:

- Detección de renglones
- Separación de caracteres
- Extracción de características
- Introducción en la red de neuronas
- Reconocimiento final del texto

3.3 ESQUEMA BASICO DE OCR CON REDES NEURONALES:

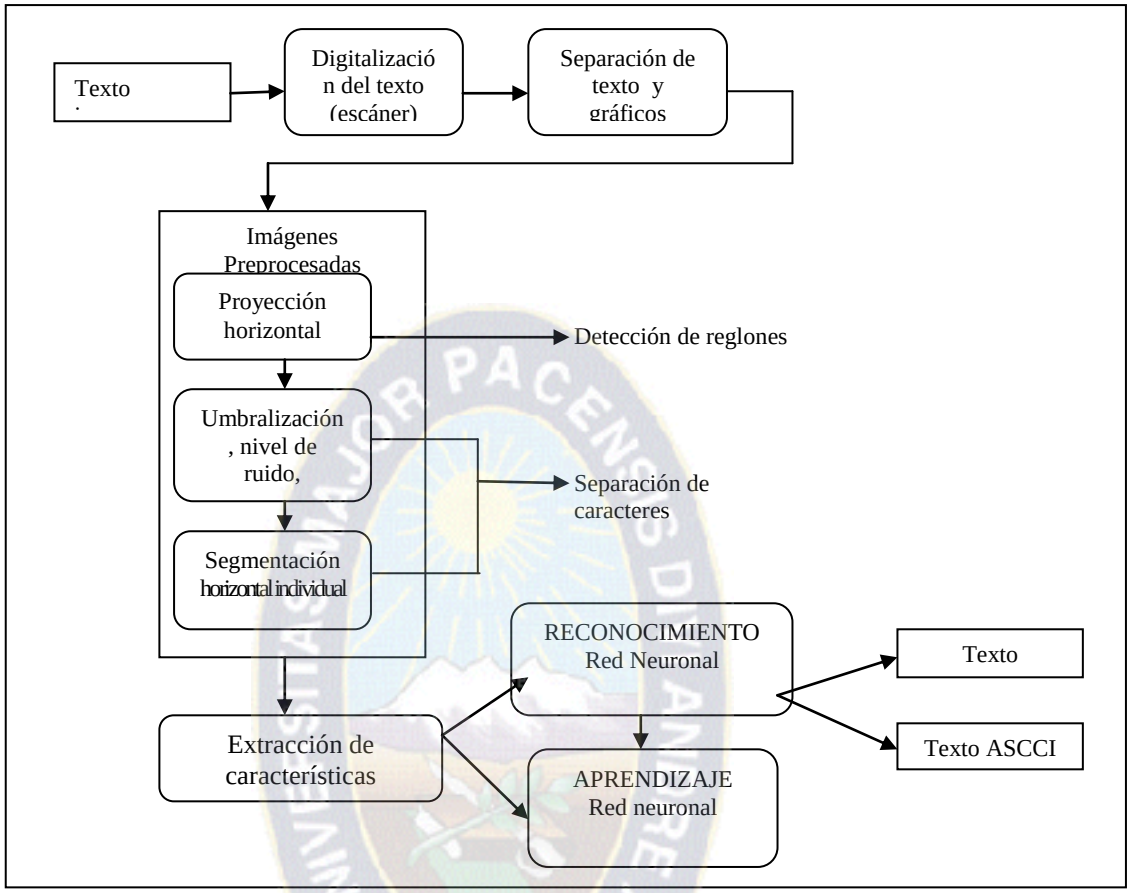


Figura 3.5

Fuente [Elaboración propia]

3.4 DESCRIPCION FORMAL DEL MODELO

En el segundo capitulo se demostró que las expresiones regulares y los autómatas finitos traductores, son un mecanismo lo suficientemente expresivo para definir el análisis léxico de lenguajes. La definición teórica precisa de los mecanismos a utilizar, se propuso la implementación un modelo de automata que consta de funciones de analisis sintactico y semantico a partir de expresiones regulares lineales. Un automata que automatice la tarea de traduccion.

Si bien en teoría queda resuelto el problema de construir este modelo traductor, hay que tener en cuenta que en la práctica es bastante difícil hacerlo. Ya que, pasar de la definición de un lenguaje por medio de expresiones regulares al autómata que lo reconozca; si bien es un proceso mecánico, la traduccion puede llegar a ser una tarea compleja. Hay que resaltar que cada lenguaje requiere la construcción de un analizador léxico que lo reconozca y traduzca.

3.4.1 GRAMATICA BASICA DEL CHINO

Para el diseño del proceso de traduccion automatica de acuerdo a nuestro limite temático. Es necesario entender la gramatica del chino, lo cual implicara una breve descripcion de la lexicologia, sintaxis y semantica de esta lengua.

En el alfabeto español tenemos 27 letras. Pero no hay letras en chino mandarín. Aquí yace el mayor error, el chino no tiene letras, ya que su idioma funciona mediante sílabas. Cada ideograma o carácter chino representa una sílaba con su respectivo tono. Por ejemplo el carácter 你 representa la sílaba nǐ (ese acento sobre la 'i' representa el tono, ya que el mandarín posee cuatro tonos y un tono neutro) y significa 你 = tú. Sin embargo, puede haber otras sílabas nǐ representadas por un carácter distinto y con un significado diferente.Existen dos reglas fundamentales para empezar a observar la lengua china:

1) La construccion de la frase simple corresponde al enunciado:
"sujeto + verbo + complemento" que se puede indicar tambien con la formula "tema + predicado + complemento".
2) El determinante (el adjetivo, por ejemplo) precede al determinante (el nombre, por ejemplo); se dirá "blanco caballo" 白马 bái mǎ, (bai maa) y no "caballo blanco".

El elemento inicial de la frase china (el "sujeto") designa la persona, la cosa, la idea, la localización, la duración, etc. Como ya lo hemos indicado, los

caracteres chinos, y las palabras que ellos utilizan para construirlas, son invariables; y que en algunos casos no hay género masculino ni femenino y los artículos no existen, mǎ 马 significa: un caballo, el caballo, unos caballos, los caballos, sin precisar.

El predicado puede ser verbal o nominal: **los verbos chinos no se conjugan;** igual que los nombres, son invariables; es el contexto (hoy, ayer, el año próximo, etc) ó el uso de ciertas partículas lo que indican como deberán estar conjugados en la traducción española.

NOTA: Recordar que trabajaremos con el pīnyīn de los caracteres chinos, que fue creado justamente con el propósito de exteriorizar el idioma en su comprensión y pronunciación, para los demás países del mundo. Y que hoy en día desde 1979, primaria y secundaria del sistema educativo chino también lo aprenden.

3.5 ANÁLISIS LÉXICO

El análisis léxico toma una cadena de caracteres y produce una cadena de nombres, palabras reservadas y signos de puntuación (a partir del alfabeto de entrada); este descarta espacios en blanco. La principal razón de separar el análisis léxico del analizador gramatical semántico reside en simplificar la complejidad de este último.

3.6 ANÁLISIS SINTÁCTICO

Estudia el orden y la relación de los caracteres chinos en una palabra que conforma la oración. A partir de la morfologia y la lexicolologia, se puede construir automáticamente un analizador sintáctico eficiente; que determina si una especificación fuente está sintácticamente bien formada. También permite identificar ambigüedades sintácticas en el proceso de construcción del

analizador sintáctico. El funcionamiento del analizador sintáctico es idéntico al del analizador sintáctico de un compilador. Obtiene una cadena de componentes léxicos del analizador léxico y comprueba si la cadena puede ser generada por la gramática del lenguaje fuente.

3.6.1 AUTOMATA ANALIZADOR SINTÁCTICO

Para nuestro caso este análisis se divide en dos etapas:

Etapa 1 en está etapa el autómata verificara que cada símbolo es parte del alfabeto de entrada con lo cual garantizara que la palabra introducida esta reconocida dentro del alfabeto pīnyīn del chino.

Lo que significa que $a_i \in$ Sa del automata; su origin es el alfabeto pinyin del chino es decir:

Sa = {á,ā,ă,à,b.c.d.é,ē,ĕ,e,f,g,h,í,ī,ĭ,l,j,k,l,……p,q,r,s,t,ú,ū,ŭ,u,v,w, x,y,z/ Á,Ā,Ă,A,B,C, D,E,F,G,……..,W,X,Y,Z },de acuerdo a la figura 3.6 siguiente donde cada $a_i \in$ Sa.

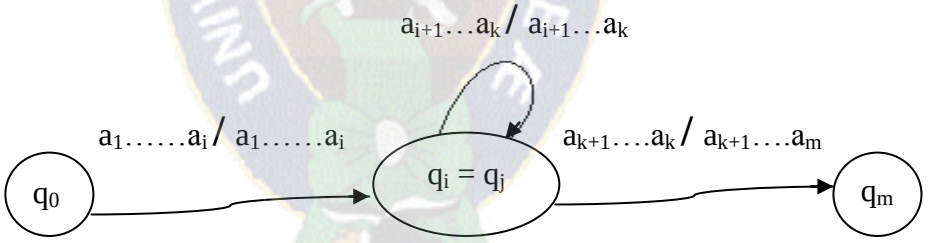


Figura 3.6

Fuente [Elaboración propia]

Etapa 2 en esta segunda etapa el autómata analizara cada palabra (**w)** escrita en chino y su correspondiente pīnyīn, para evidenciar si está escrito de correcta forma el pīnyīn del idioma.

Este automata recibe como entrada la palabra pīnyīn, dividida en caracteres en un estado q_0**,** cuando encuentra en espacio vacio pasa a un estado q_1, si el carácter siguiente es un carácter sintactico complementario para completar la significación pasa al estado q_2, esperando vacio o un carácter para pasar al

estado q_3 si el carácter que sigue no fuera el complementario y es espacio vacio entonces vuelve al estado q_0.

SIX = < Q, Sa, AS, Re, q_0, F> donde:

1) Q es el conjunto de estados.

2) Sa={á,ā,ă,à,b.c.d.é,ē,ĕ,e,f,g,h,í,ī,ĭ,l,j,k,l,……p,q,r,s,t,ú,ū,ŭ,u,v,w, x,y,z/ Á,Ā,Ă,A,B,C, D,E,F,G,……..,W,X,Y,Z } alfabeto de entrada.

3) q_0 estado inicial, $q_0 \in Q$.

4) AS: QxSa→Q la función que realiza la categorización sintáctica en un estado.

5) Re: Q x Sa → Sb es la función de respuesta que considera la información válida para poder procesarla.

6) F: {q_j} conjunto de estados finales o de aceptación.

La máquina se encuentra en el estado q_0. Posteriormente, cuando la máquina se encuentra en un estado $q_i \in Q$ y recibe un símbolo **e** perteneciente al alfabeto de entrada Sa, entonces emite un símbolo de salida s = Re(q,e) y transita al nuevo estado r = AS(q,e). Ejemplo:

Cararts. Chinos	**pīnyīn**	**pronunc. prox.**
自马	bái mă	(bai maa)

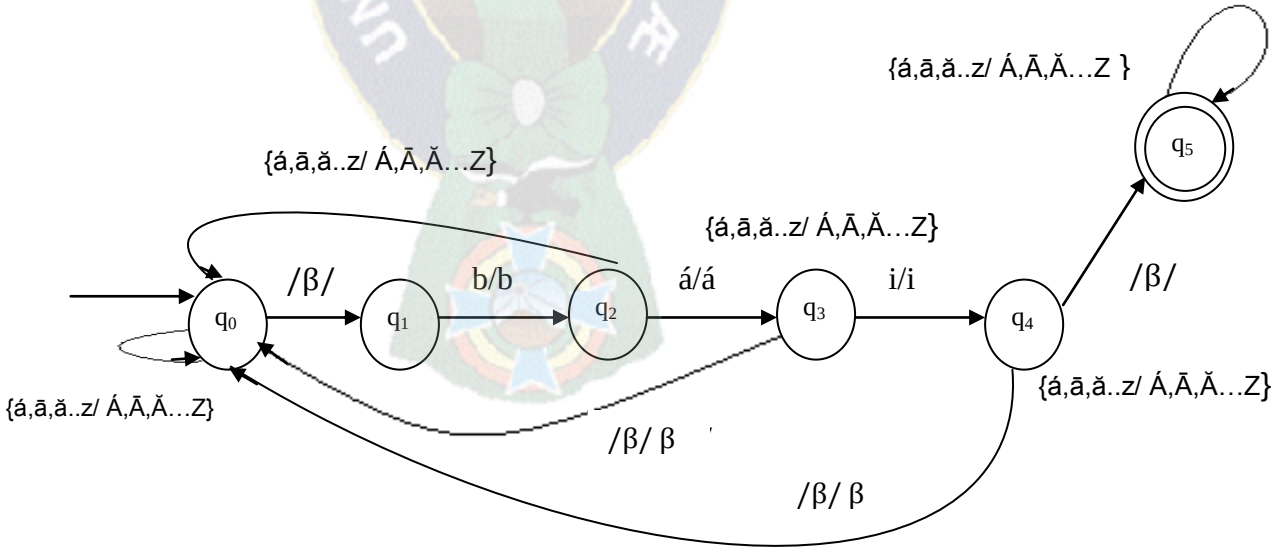


Figura 3.7: Verificación de **"bái"**

Fuente [Elaboración propia]

El automata recibe como entrada el alfabeto de entrada, en un estado q_0 cuando encuentra un espacio vacio β pasa a un estado q_1, si el carácter siguiente es “b” pasa al estado q_2 , esperando la letra “á” si el carácter siguiente no fuera “á” entonces vuelve al estado q_0, si la letra es “á” pasa al estado q_3, esperando la tera “i”,si el carácter siguiente no fuera “i” entonces vuelve al estado q_0, si la letra es “i” pasa al estado q_4.

Si se encuentra en el estado q_4 espera un espacio vacio, para pasar al estado de satisfación q_5, caso contrario vuelve al estado q_0.

El estado q_5 verifica que la frase presente **“bái”** es de aceptación, para poder utilizar una de las reglas de traducción.

Ahora trabajaremos con la siguiente palabra pīnyīn: **mā**

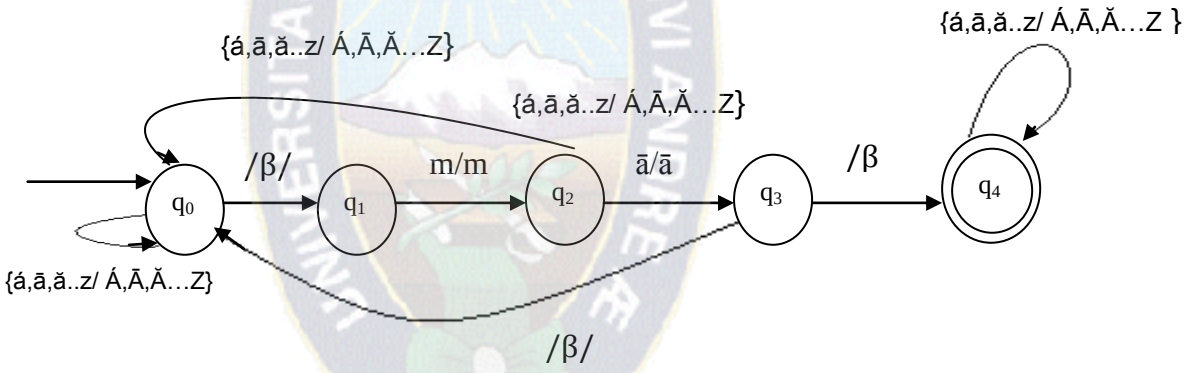


Figura 3.8: Verificación de **“mā”**

Fuente [Elaboración propia]

El automata recibe como entrada el alfabeto de entrada, en un estado q_0 cuando encuentra un espacio vacio β pasa a un estado q_1, si el carácter siguiente es “m” pasa al estado q_2 , esperando la letra “á” si el carácter siguiente no fuera “ā” entonces vuelve al estado q_0, si la letra es “ā” pasa al estado q_3, si se encuentra en el estado q_3 espera un espacio vacio, para pasar al estado de satisfación q_4, caso contrario vuelve al estado q_0.

El estado q_4 verifica que la frase presente **“mā”** es de aceptación, para poder utilizar una de las reglas de traducción.

3.7 ANÁLISIS SEMÁNTICO

Permite conocer el significado de los signos lingüísticos y de sus combinaciones en una oración, por lo tanto, está vinculada al significado, sentido e interpretación de palabras, expresiones o símbolos.

La semántica estudia la codificación del significado en el contexto de las expresiones lingüísticas.La **denotación** (la relación entre un palabra y aquello a lo que se refiere) y la **connotación** (la relación entre una palabra y su significado de acuerdo a ciertas experiencias y al contexto).

El estudio del **referente** (aquello que la palabra denota, como un nombre propio o un sustantivo común) y del **sentido** (la imagen mental que crea el referente)

Este automata recibe como entrada la frase dividida en palabras (lexemas) en un estado q_0**,** cuando encuentra en espacio vacio pasa a un estado q_1, si la palabra siguiente completa la frase semanticamente pasa al estado q_2, esperando vacio o la palabra para completar la significación pasa al estado q_3, si la plabra que sigue no fuera el complementario o espacio vacio entonces vuelve al estado q_0.

3.7.1 AUTOMATA ANALIZADOR SEMÁNTICO GRAMATICAL:

Que hará uso de las reglas gramaticales.

SEM = < Q, Sb, ASe, L, q_0, F>

donde:

1) Q es el conjunto de estados.

2) Sa

={á,ā,ă,à,b.c.d.é,ē,ĕ,e,f,g,h,í,ī,ĭ,l,j,k,l,……p,q,r,s,t,ú,ū,ŭ,u,v,w,

x,y,z/ Á,Ā,Ă,A,B,C, D,E,F,G,……..,W,X,Y,Z } alfabeto de entrada.

3) **Sb** es el conjunto de palabras a ser analizadas, Sb ⊆ Sa

4) q_0 estado inicial, $q_0 \in Q$.

5) ASe: Q x Sb → Q es la función que permite analizar el contexto semántico de expresiones o símbolos.

6) L: Q x Sb → Sc es la función de respuesta que considera la información válida para poder procesarla.

7) F: { } conjunto de estados finales o de aceptación.

Ejemplo:	**PĪNYĪN**	**Sujeto**	**Verbo**	**Adjetivo**	**Nomb/Sust**
	wǒ xúe zhōng wén	wǒ	xúe	zhōng	wén
	(yo estudiar china lengua)				

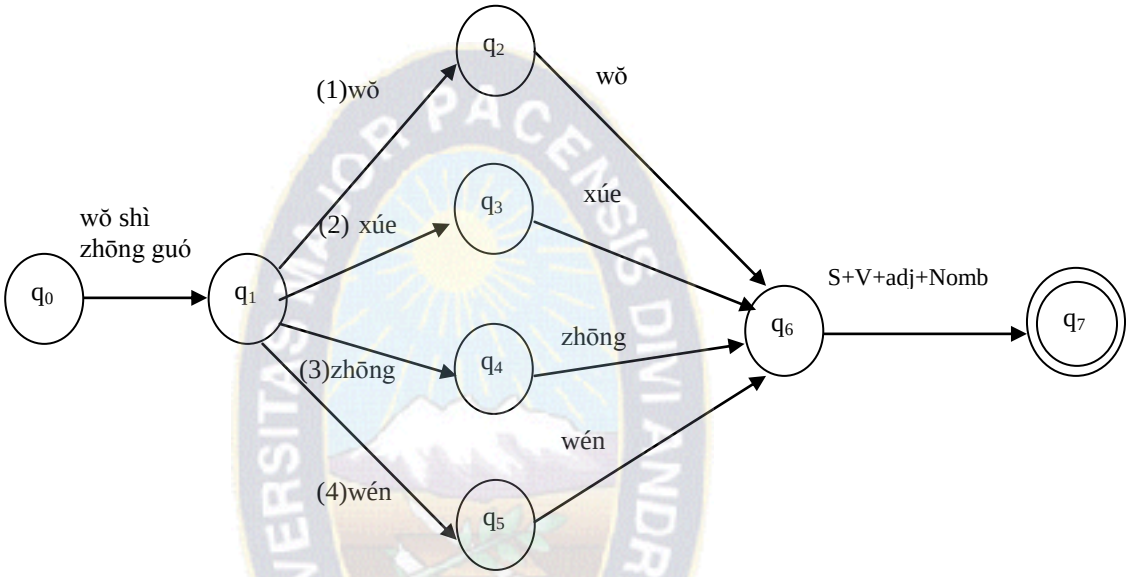


Figura 3.9: Verificación semantica

Fuente [Elaboración propia]

Ejemplo: maǐ , este carácter solo; significa “comprar”, “sobonar”, etc.

maì , este carácter solo; significa “vender”, hacer alarde de, etc.

(*) Los dos caracteres juntos: maǐ maì pronun.aprox.*(mail mail)* significa “comercio”,”negocio”, “trocar”,”comerciar”, etc.

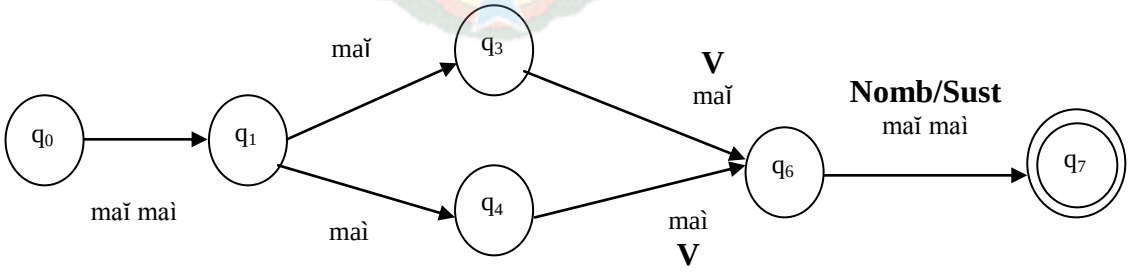


Figura 3.10: Verificación semantica

Fuente [Elaboración propia]

Ejemplo:

Caract. chinos PĪNYĪN Sujeto verbo adjetivo de ∈ Nomb/Sust.

Wǒ shì xī bàn yá rén Wǒ shì xī bàn yá rén

(yo soy españa hombre)

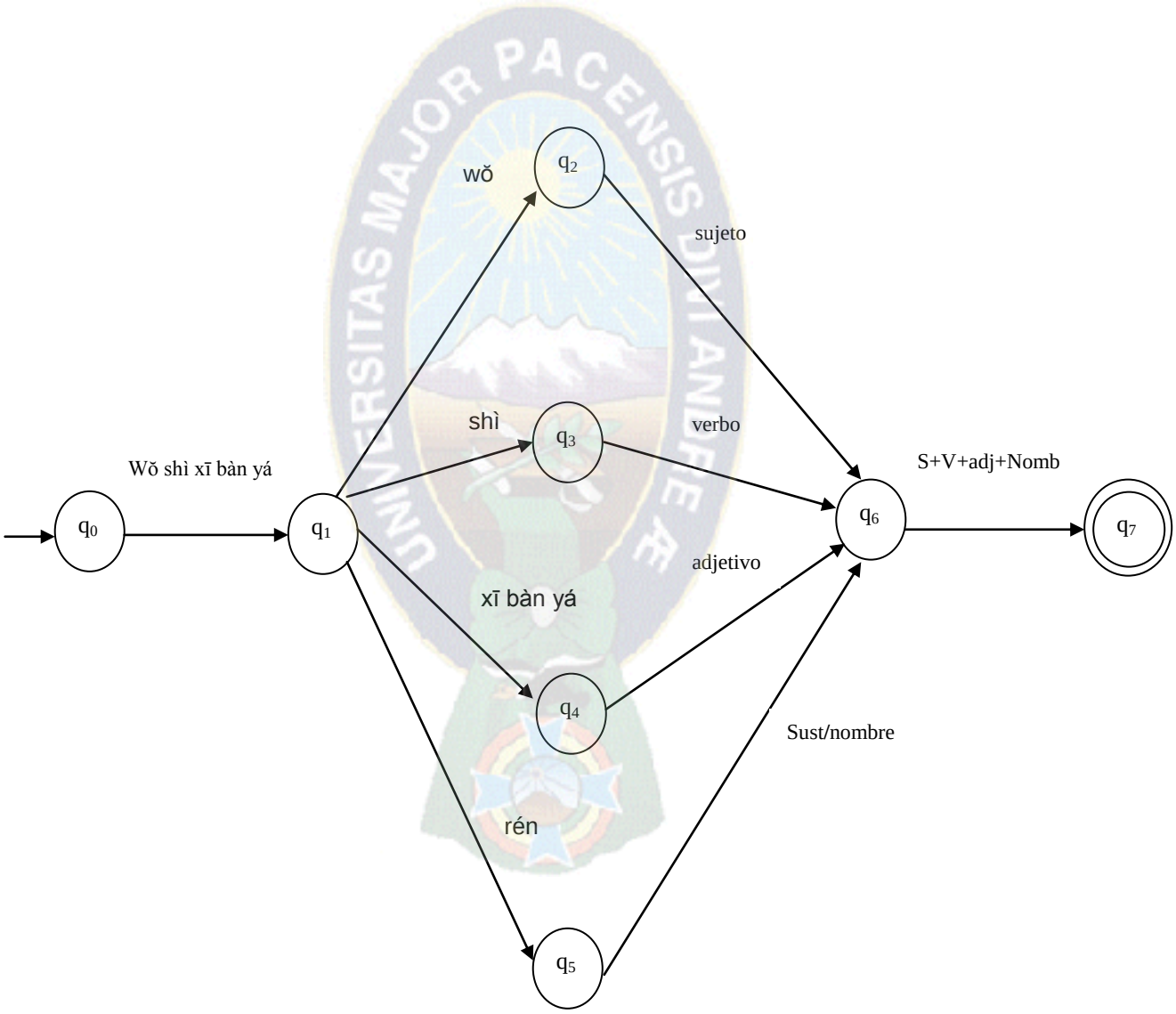


Figura 3.11: Verificación semantica / gramatical

Fuente [Elaboración propia]

3.8 MODELO AUTÓMATA FINITO TRADUCTOR

Caso: (Chino a Español)

Un autómata finito traductor MT es simplemente un autómata finito que se define como una 7-upla:

MT = <Q, ∑, δ, q_0, F, S, t>.

Donde:

Q: Conjunto finito de *estados*,

∑*: Alfabeto* o conjunto finito de símbolos de entrada **(caracteres chinos** Sc **)**

δ: Es la *función de transición de estados* definida **δ: Q x ∑ →*Q***

q_0: *Estado inicial* **q_o ∈ ∑.**

F: Conjunto de *estados finales o estados de aceptación.* **F⊆Q**.

***S:** Alfabeto* o conjunto finito de símbolos de salida**(palabras en español)**

T: Es la *función de traducción* definida **T: Q x ∑ → *S****

Ambas funciones **δ:** Q **x** ∑ → *Q* y **T:** Q **x** ∑ →*S** están definidas sobre Q **x** ∑ si existen $\delta(e_i,a) = e_j$.

NOTA:

Para una mejor comprension del caso; en la función **δ** se denotara como **x** al alfabeto de entrada a reconocer (leng. chino) y en la función **T** se denotara **a** al alfabeto de salida resultante (leng. español)

Si existen **δ(**qi , ***x*) =** *q*j y **T (**qi , ***x*) =** *a*

donde qi, qj **∈** Q; *x* **∈** ∑*; a* **∈** S* en el diagrama de transición de estados el valor de la traducción ***a*** se agrega sobre los arcos.

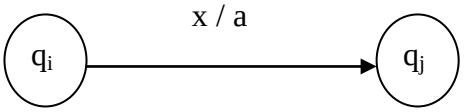


Figura 3.12
Fuente: [Elaboración propia]

3.8.1 FUNCIÓN DE TRADUCCIÓN PARA CADENAS

La extensión de la función de traducción **T*: Q x ∑* → S***, tal que **T*(qi ,ω)** es la cadena que traducirá el autómata, luego de leer la cadena **ω** en la cinta de entrada y comenzando en el estado qi, se define como:

1) t*(qi, λ) = λ

2) $t^*(qi, \omega a) = t^*(qi, \omega).\ t(\delta^*(q_i, \omega), a)$ donde $a \in \sum$, $\omega \in \sum^*$, q_0 ,qi ∈ Q.

La diferencia entre **t** y **t*** es que **t** se define desde un estado y un símbolo del alfabeto, y **t*** se define desde un estado y una cadena de símbolos.

3.8.2 Traducción:

El autómata solo define la traducción, si el autómata finito reconocedor subyacente "acepta" la cadena. Es decir, la traducción **T(ω):** $\sum^* \rightarrow S^*$ asociada a M_T está definida como: **T(ω)=** $t^*(q_0, \omega) \iff \delta^*(q_0, \omega) \in F$ donde $\omega \in \sum^*$

NOTA: La traduccion se hara utilizando las reglas gramaticales del español; en donde el adjetivo va despues del nombre o sustantivo.

Caract. chinos	Pīnyīn	Traduccion/semantica	Interpretación
	Wǒ shì xī bàn yá rén	Yo soy hombre españa (yo soy españa hombre)	Yo soy español

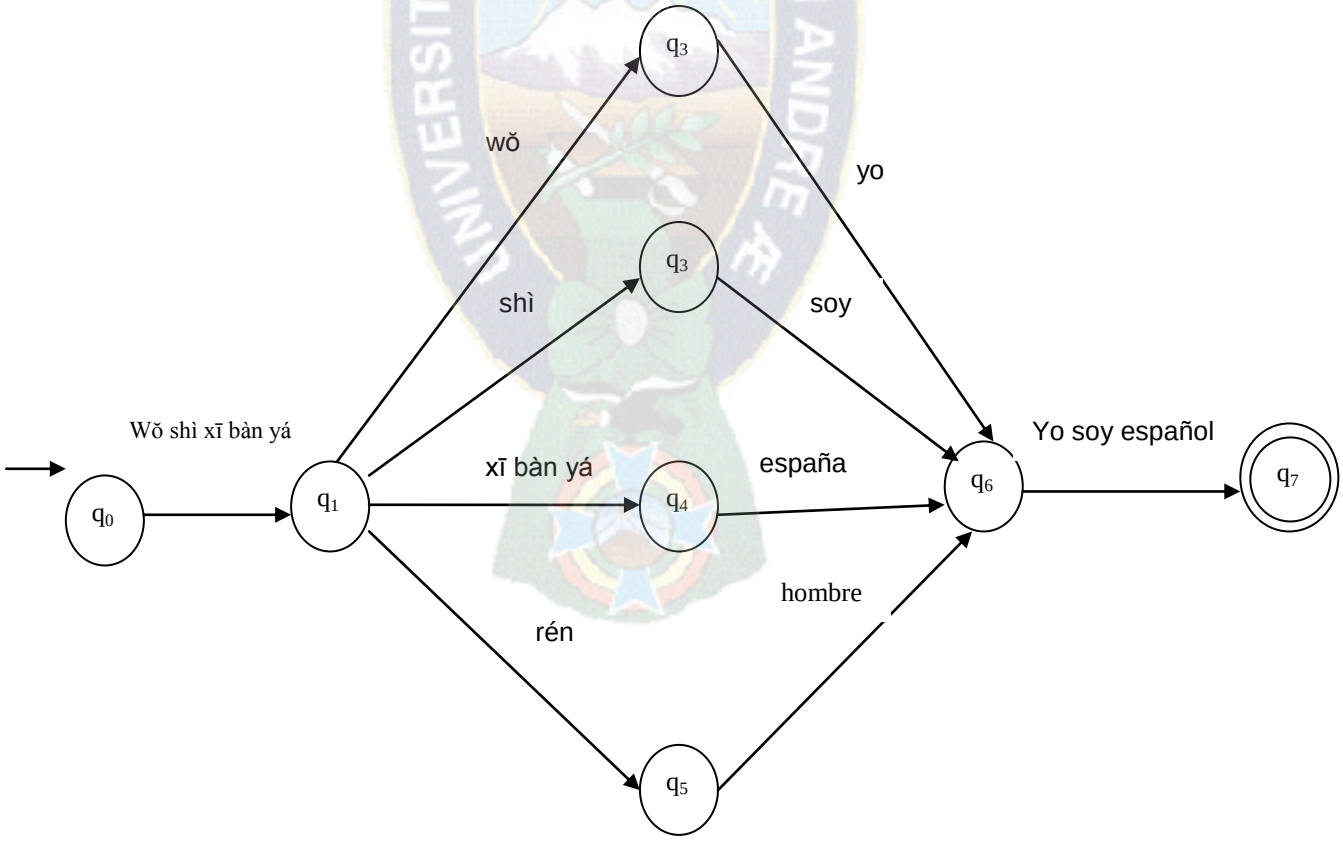


Figura 3.13: Automata aplicando regla de traducción

Fuente [Elaboración propia]

3.9 REGLAS GRAMATICALES DEL CHINO (algunas nociones)

El predicado puede ser verbal o nomial ; los verbos chinos no se conjugan; igual que los nombres, son invariables: es el contexto (hoy, ayer, el año próximo, etc…) ó el uso de ciertas partículas lo que indican como deberán estar conjugados en la traducción española.

Ideograma	Significado	Pronunciación (Pinyin)	Ejemplo
是	Ser	Shì	我**是**工人。 (Yo **soy** trabajador.)
在	Estar/Encontrarse/Existir	Zài	我**在**酒吧。 (Yo **estoy** el bar.)
買	Comprar	Mǎi	你要不要**買**一杯。 (¿Quieres **comprar** una taza?)
賣	Vender	Mài	你要不要**賣**一杯。 (¿Quieres **vender** una taza?)
喝	Beber	Hē	他**们在喝**茶。 (Ellos están **bebiendo** té.)
吃	Comer	Chī	我在**吃饭**。 (Yo estoy **comiendo**).
写	Escribir	Xiě	我在**写**字**汉语**。 (Yo estoy **escribiendo** en chino).

Algunos verbos chinos:

ESPAÑOL	CHINO	LOGOGRAMA
aconsejar	jiàn yì	建议
amar	aì	爱
apoyar - recostar	kào	靠
bailar	Tiàowǔ	跳舞
beber	hē - yǐn	喝 - 饮
buscar	zhǎo	找
caminar	zǒu	走
comer	chī	吃
contestar	huí dá	回答
correr	pǎo - ben pǎo	跑 - 奔跑
dar	geǐ	给
decir	shuō	说
descansar	xiūxī	休息
dormir	shuìjiào	睡觉

ejercitar	duànliàn	**锻炼**
entrar - avanzar	jìng	**进**
escuchar - oir	Tīng	听
estudiar	Xuéxí	**学习**
ir	qù	去
leer	kàn	看
llegar -arrivar	dàodá	**到达**
mirar, ver	kàn	看
mover	dòng	**动**
oir	wén	**闻**
partir - salir	lí - líkāi	**离 - 离开**
pasear	Sànbù	**散步**
recibir	shōudào	收到
saber	zhidao	知道
salir	chū	出

saltar	tiào	跳
sentar	zuò	坐
tener	Yǒu	有
trabajar	Gōngzuò	工作
venir	lái	来
viajar	fēi	飞

El adverbio es la clase de palabra que actúa como núcleo del sintagma adverbial. Modifica al verbo, a un adjetivo o a otro adverbio.

Esta mañana	jīn zǎo - 今早
semana próxima	xià zhōu - 下周
ya	yǐ jīng - 已经
recientemente	zuì jìn - 最近
últimamente	jìn lái - 近来
pronto	bù jiǔ - 不久
inmediatamente	lì jí - 立即
aún	réng - 仍
todavía	hái;huán, réng rán, shàng - 还，仍然，尚
hace…	yǐ qián - 以前
adverbios de lugar	dì diǎn fù cí - 地点副词

aquí	zhè lǐ - **这里**
ahí	nà lǐ - **那里**
por ahí	zài nà biān - **在那边**
en todas partes	dào chù - **到处**
dondequiera	rèn hé de fāng - **任何地方**
en ninguna parte	wú chù - **无处**
a casa	zài jiā, huí jiā - **在家，回家**
lejos	lí kāi, zǒu kāi - **离开，走开**
fuera	chū wài, zài wài, xiàng wài, lí qù - **出外，在外，向外，离去**
adverbios de manera	fāng shì fù cí - **方式副词**
muy	hěn, fēi cháng - **很，非常**
bastante	xiāng dāng - **相当**
mucho	xiāng dāng, pō - **相当，颇**
realmente	zhēn zhèng de, què shí de - **真正地，确实地**
rápido	kuài sù de, xùn sù de - **快速地，迅速地**
bien	hǎo;hào - **好**
duro	nǔ lì dì;de - **努力地**
rápidamente	hěn kuài - **很快**
despacio	màn màn de - **慢慢地**
cuidadosamente	xiǎo xīn dì, rèn zhēn de - **小心地，认真地**

apenas	jī hū bù - 几乎不
difícilmente	jī hū bù - 几乎不
en su mayoría	dà bù fēn, zhǔ yào de - 大部分， 主要地
casi	jī hū - 几乎
absolutamente	jué duì - **绝对**
juntos	yī qǐ - 一起
solo	dān;chán - **单**
adverbios de frecuencia	pín dù fù cí - **频度副词**
siempre	zǒng shì - **总**是
frecuentemente	cháng cháng, pín fán de, jīng cháng de - **常常，频繁地，经常**地
por lo general	tōng cháng, píng cháng - 通常，平常
a veces	yǒu shí - 有**时**
de vez en cuando	ǒu ěr - **偶尔**
raramente	hěn shǎo - 很少
rara vez	hěn shǎo - 很少
nunca	cóng bù - 从不

Los adjetivos en chino

El adjetivo es la palabra que acompaña al sustantivo o nombre para determinarlo o calificarlo; expresa características o propiedades atribuidas a un sustantivo, ya sean concretas (el libro **verde**, el libro **grande**).

fácil	róng yì - 容易
vacío	kōng de - 空的
caro	áng guì - 昂贵
rápido	kuài - 快
extranjero	guó wài - 国外
completo	mǎn - 满
bueno	hǎo;hào - 好
duro	yìng - 硬
pesado	zhòng;chóng - 重
barato	pián yí de - 便宜的
ligero	qīng - 轻
local	dāng dì - 当地
nuevo	xīn - 新
ruidoso	cáo zá - 嘈杂
viejo	lǎo - 老
poderoso	qiáng dà - 强大
tranquilo	ān jìng - 安静
correcto	zhèng què - 正确
lento	huǎn màn - 缓慢
suave	ruǎn - 软
muy	hěn - 很
débil	ruò - 弱

mojado cháo shī - **潮湿**

mal cuò wù - **错误**

joven nián qīng - **年轻**

cantidades shù liàng - **数量**

pocos hěn shǎo - **很少**

poco hěn shǎo - **很少**

muchos xǔ duō - **许多**

mucho xǔ duō - **许多**

parte bù fēn - **部分**

algunos yī xiē - **一些**

unos pocos jǐ gě - **几个**

Fuente[Beijing language and cultura,University press]

3.10 REGLAS GRAMATICALES DEL ESPAÑOL (algunas nociones)

Las palabras gramaticalmente, son conocidas tambien como partes de la oración, estas son nueve: nombre sustantivo, adjetivo, verbo, adverbio, pronombre, articulo, preposición, conjunción e intrerjección.

La gramatica estructural agrupa a las palabras en seis clases: sustantivo, verbo, adverbio, adjetivo, preposición y conjunción. Descritas morfologicamente:

Sustantivo: Es una categoria gramaticall que admite morfemas de número, genero, determinanates y afijos, sustantivos personales, conocidos como pronombres personales tónicos

yo	ella	vosotros(as)
tú	ello	ellos/ellas
el	nosotros(as)	ustedes

Adjetivo: admite morfemas de número, de genero y morfemas facultativos, llamados tambien afijos. Se divide en adjetivos calificativos, adjetivos determinativos son los pronomb. demostrativos, posesivos, indefinidos, numerales y relativos

Grado positivo	Grado comparativo			Grado superior	
Cualidad	Igualadad	Superid.	Inferiord.	Absoluto	Relativo
Es bueno	Es bueno igual	Es más bueno	Es menos bueno	Es buenisimo	Es muy bueno

Pronombre: es una unidad lingüística que puede cumplir las funciones de sustantivo, adjetivo, determinativoy eventualmente las funciones de adverbio.

Personales tónicos	Personales tónicos	Demostrativos singular	Posesivos tónico mascul.	Indefinidos sustantivo	Numerales cardinales	Relativo. pronomb.	Interrogativ exclamativo sustant.
Yo,tu	Me, te	Este	Mio, tuyo	Alguien,algo	Un/o/a	Que	Qué?
El, ella,ello	Se, lo,la	Ese	Suyo	Nadie, nada	Dos	Quien	Quién, quiénes
Nosotros(as)	Nos	Aquel	Mios	Quienquiera	Tres	Quienes	Cuál, cuáles
Vosotros(as)	Os	Esta, esa	Tuyos	Cualquiera	Cien	Cual	Qué
Ellos, ellas	Se les	Aquella	Suyos	Todo, mismo,uno	Mil	cuales	Cúyo
					Un millon		cúya

El articulo:es una categoria gramatical cuya unica ffunción es determinar la sustantivo, puede sustantivar otras categorias pero no sustantivos, como adjetivos y verbos.		
Masculino	el	Los
Femenino	la	Las
Neutro ó articulo invariable	Lo	

El verbo: es la categoria lingüística más importante ya que en ella se fundamenta la construccion de la oración; esta categoria gramatical expresa la acción de un fenomeno, el proceso de realización de una actividad. Puede adquirir dos formas: predicada y otra apredicada		
Formas apredicadas		
Infinitivo	-ar, -er, -ir	Cant-ar
Gerundio	-ando, -endo, -yendo	Cant-ando
Participio	-do, -da	Canta-do
	-to, -ta	Abier-to
	-cho, -cha	Satisfe-cho
	-so, -sa	Impre-so

El adverbio: es el elemento gramatical que expresa circunstancias varias de la acción verbal. Sintacticamente actua como modificador del verbo y desempeña funciones de complemento circunstancial en la oración. Los adverbios son de: modo, tiempo, lugar, cantidad, afirmación, negación, duda, posibilidad.	
Adv. de tiempo	Ayer, hoy, mañana, siempre, tarde
Adv. de lugar	Ahí, aquí, allá, cerca, atrás
Adv. de modo	Así, despacio, bien, mal
Adv. de cantidad	Mucho, demasiado, más, muy
Adv. de afirmación	Sí, efectivamente, evidentemente
Adv. de negación	No, nunca, jamás
Adv. de duda	Quizás, talvez, posiblemente, probablemnete
Adv. de deceo	ojalá

La preposición: La preposición es una categoria gramatical que sirve para relacionar y conectar palabras, sintagmas y frases con otras de su misma naturaleza
A, ante,bajo, con, contra, de ,desde, en entre, hacia, hasta, para, por, según, sin, so, sobre, tras, cabe(esta ultima en desuso).

La conjunción: Las conjunciónes son palabras de carácter invariable, con valor puramente gramatical cuya unica función en la lengua es enlazar unidades linguisticas, por lo general proposiciones u oraciones.	
Copulativa	y. e. ni. que
Disyuntivas	o, u
Adversativas	Mas, pero, sino, sin embargo
Distribuitivas	Ya…, ya, bien…bien.
Explicativas	Esto es, es decir, o sea

3.11 ETAPA DE INGENIERIA DE TRADUCCION AUTOMATICA

La implementacion del traductor automatico seguira la arquitectura de transferencia, que traducira automatIcamente sintagmas nominales del idioma chino al idioma español, la cual se realizara en tres fases: analisis, traducción por transferencia y generación.

El nivel de transferencia sera lexico sintactico y semantico gramatical.

Transferencia lexica: permitira la busqueda del termino equivalente en la lengua meta la cual se realizara a partir de la información contenida en el diccionario contenido en una base de datos interno.

Transferencia sintactica: el analisis de la oración de la lengua origen se transformara en la generación equivalente para la oración de la lengua meta.

3.11.1 etapa de implementación del traductor y reglas gramaticales

Las entradas del traductor se realizara mediante archivos ocr; entradas en el diccionario monolingue chino, para que el traductor pueda analizar (“entender”) la palabra cuando, la encuentre en un texto, y la puede generar al traducir esta palabra al español.

Entrada en el traductor bilingüe; para que el sistema pueda traducir esta palabra del idioma chino al idioma español.

Entrada en el traductor monolingue del español; para que el traductor pueda analizar (“entender”) la palabra cuando la encuentre en un texto y la pueda generar al traducir esta palabra al idioma español.

3.11.2 IMPLEMENTACIÓN DE LAS REGLAS GRAMATICALES

El módulo de transferencia estructural (generado a paritir del fichero de reglas de transferencia estructural) llamará durante el procesamiento, al módulo de transferencia lexica (generado a partir del traductor bilingue) para determinar los equivalentes en el lenguaje meta de las formas lexicas.

Las reglas de transferencia se implementaran en archivos ocr. las reglas tendrán un patrón y una acción. El patron indicara que secuencias de formas léxicas tienen que seer detectadas y procesadas. La acción describirá las verificaciones y transformaciones que deben realizarse en ellas.

3.12 DESARROLLO DE TRADUCCIÓN

El desarrollo de este prototipo tiene como objeto brindar una idea general de los automatas para la traducción de una frase (sintagma) el cual esta implementado en Visual .Net y la base de datos sql, dando la posibilidad de implementar programas que permitan mostrar una mejor interfaz entre el usuario y el programa.

Implementación:

La traducción sintactica y semántica del idioma chino al español utilizando automatas, lleva consigo las cualidades de interfaz comparables con algunos traductores, para nuestro caso en el idioma chino, ya sabemos que la traducción no es directa qu edebemos hacer uso del correspontiente pīnyīn para luego poder traducir al español, entonces nuestro prototipo presenta tres ventanas textuales.

Para la traducción se captura los caracteres del documento digitalizado, en la primera ventana superior luego, presione clic sobre el boton de traducir , esto desplegara en la ventana siguiente su correspondiente pinyin y en la ventana inferior la frase traducida al español.

Pantalla inicio:

Figura 3.14

Fuente [Elaboración propia]

Para salir del programa presione clic sobre el boton salirdel menú de arriba ó si decea guardar el documento, clic en guardar doc. Traducido y si desea aprender su correspondiente pinyin, clic en guardar doc. con pīnyīn.

La presentación de la experimentación empezara con las entradas de palabras, frases, oraciones cortas, se analizaran automaticamente y se obtendran como resultado, el equivalente de la frase traducida en latercera ventana de la pantalla, y si fuese necesario en cualquier momento se realizara la busqueda de una palabra en el diccionario de la base de datos contenida en el sistema para posibilitar un conocimiento mas profundo.

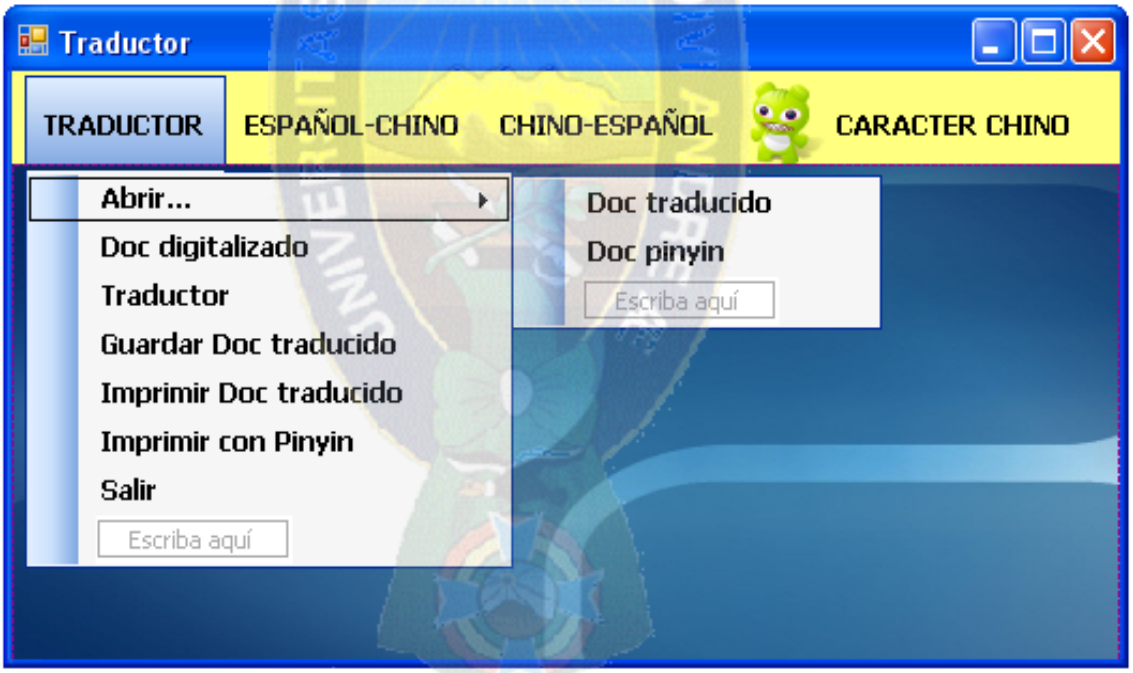


Figura 3.15

Fuente [Elaboración propia]

Primero: descompondrá la frase de entrada en palabras y lo contendrá en una matriz.

Segundo: clasificara la palabra según su categoría gramatical: sujeto, verbo adjetivo, adverbio, conector, etc.

Tercero: traducirá palabra por palabra, en el caso del chino logograma por logograma.

Cuarto: ordena en función de la ubicación del sujeto, verbo, conector y complemento, entonces el autómata decidirá corresponder con su correspondiente en el idioma español, que está identificado en la base de datos.

Quinto: simplifica y completa las palabras de acuerdo al idioma que está traduciendo.

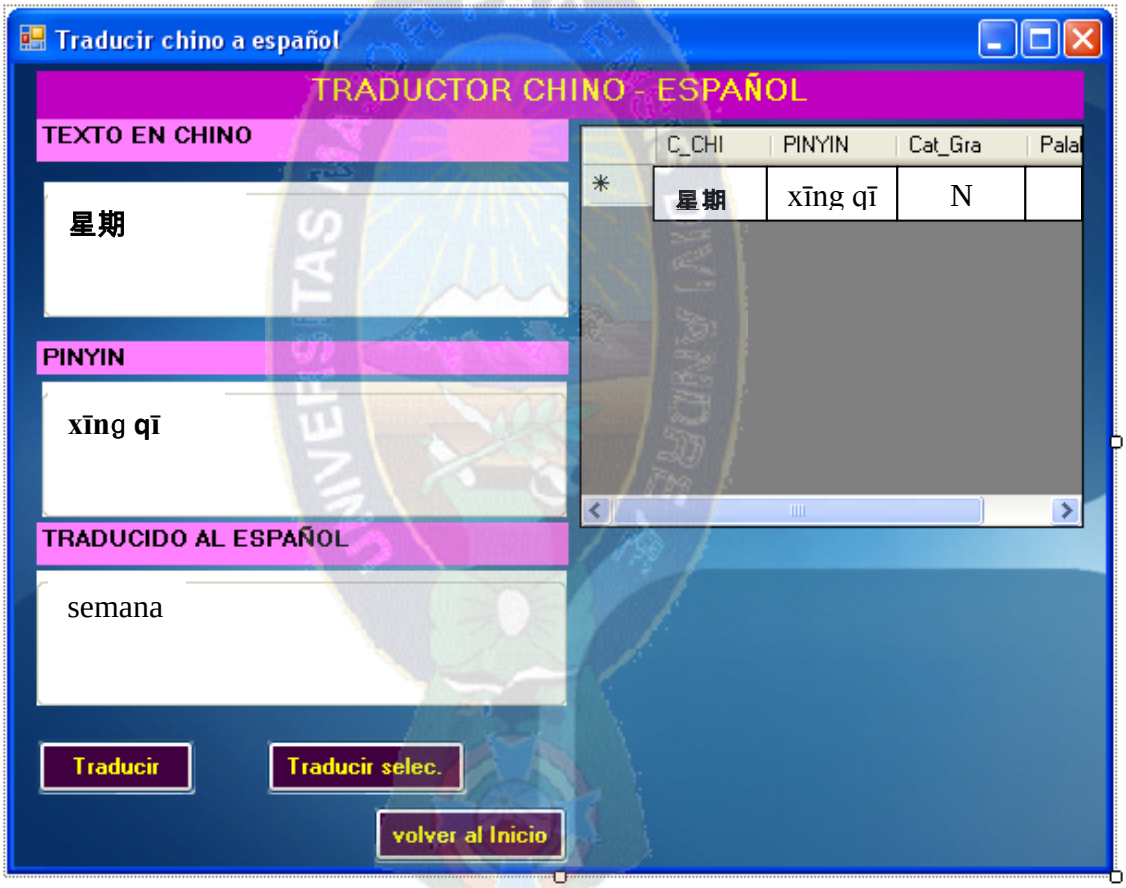


Figura 3.16

Fuente [Elaboración propia]

Traductor de frases, puede traducir en forma general pero no muy precisa. Entonces luego deberá ser interpretada por el usuario de acuerdo a contexto.

Traductor de carácter chino: contiene una traducción más exacta.

El traductor: es el que traduce las palabras, frases y documentos.

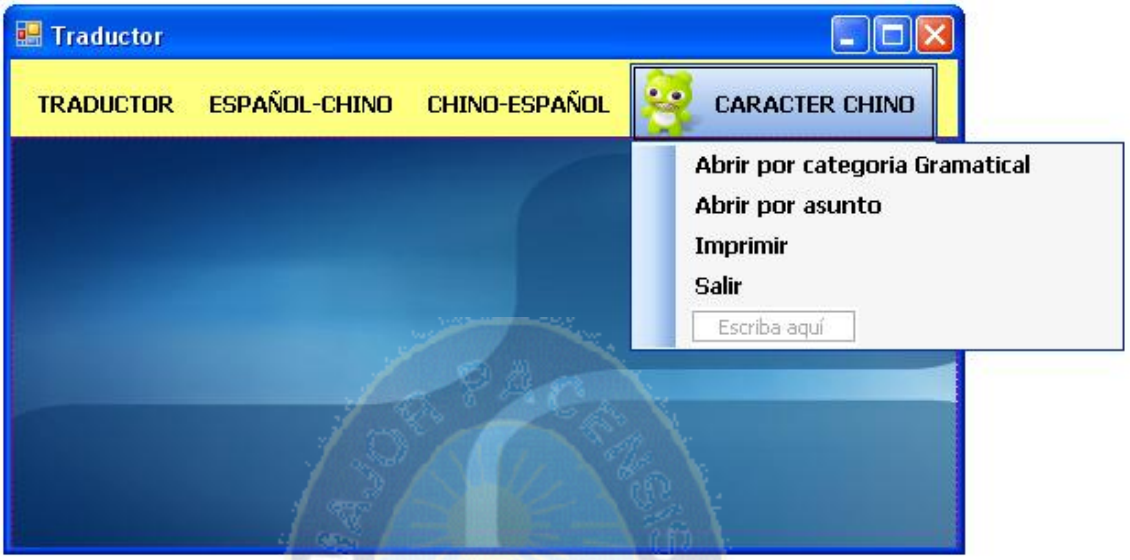


Figura 3.17

Fuente [Elaboración propia]

Salir: es para salir del programa.

Imprimir doc. traducido: imprime el documento ya traducido que está en esa pantalla.

Imprimir doc. con pīnyīn. Imprime el documento con su correspondiente pīnyīn y su traducción en el idioma español.

Abrir doc. traducido : abre el documento en chino que ya fue traducido al idioma español.

Abrir doc. pīnyīn: abre el doc. con su correspondiente pīnyīn y su traducción en el idioma español.

Guardar proyecto chino-español: nos permite guardar ambos proyectos de la traducción.

Traducir selección: traduce el texto seleccionado por el usuario.

Revisión ortográfica china: revisa las palabras según el diccionario en chino.

Revisión ortográfica español: revisa las palabras según el diccionario en español.

Diccionario chino: contiene todas la palabras del vocabulario de acuerdo a nuestro limite temático que es usado por el traductor en idioma chino.

Diccionario español: contiene todas la palabras del vocabulario de acuerdo a nuestro limite temático que es usado por el traductor.
La ayuda está compuesta por 3 documentos.
Gramática china: es un resumen de las reglas gramaticales del idioma chino usadas en este sistema.
Gramática española: es un resumen de las reglas gramaticales del idioma español usadas en este sistema.
Manual de usuario: es la descripción de uso del sistema para el usuario.

El traductor permite copiar un texto en la parte superior y luego lo traduce según las normas gramaticales y coloca la traducción en la parte inferior.

3.13 RELACIÓN DEL MODELO

Permite una segmentación adecuada, desambiguación y cobertura de la gramática.
En este sentido se tomó en cuenta dos aspectos para el tratamiento de la frase y su posterior interpretación.
El uso de autómatas para la verificación de la frase y posterior uso de las reglas de traducción, con el manejo adecuado de los cuatros tonos del pīnyīn que permite traducir de manera correcta.
El uso de diccionarios contenidos en una base de datos del sistema permite al usuario reducir en gran manera el error de traducción que pudiera existir con el prototipo verificando las palabras mal traducidas en el diccionario de nuestra base de datos, que al utilizar diccionario normal.

3.14 MODELO DE PROTOTIPO

En contraste con la Ingeniería de Software de la década de los 70, que dio respuesta a proyectos grandes pero con requisitos estables, la Ingeniería de Software de los 80 reaccionó a las complicaciones resultantes de encontrarse

con requisitos poco claros y dinámicos, dando lugar a la “construcción de prototipos”. El modelo de ciclo de vida de prototipos fue propuesto por Gomaa en 1984.

Un prototipo es un mecanismo para identificar los requisitos del software. La construcción de prototipos es un proceso que facilita al ingeniero de software el desarrollo de la aplicación. El prototipo suele tomar una de las tres formas siguientes:

• Un modelo en papel o en computadora que describe la interacción hombre-máquina, de forma que facilite al usuario la comprensión de su funcionamiento. Por ejemplo, si el sistema a construir es un cajero automático, se puede hacer un programa que simule la interacción del usuario con el cajero sin que el programa esté conectado a ninguna base de datos real ni se despache dinero. De esta manera el cliente puede hacerse a la idea de cómo va a funcionar el sistema final sin tener que construirlo, y así discutirlo con el ingeniero de software. Naturalmente, en un prototipo no se simularán todas las funcionalidades del sistema pero, si es necesario, se podrán construir otros a medida que la aplicación se vaya desarrollando (ver más abajo cuáles son las etapas para su utilización)

• Un modelo que implementa una función requerida importante. Es el mismo caso que anteriormente pero sin centrarse en la interacción hombre-máquina. Por ejemplo, el modelo podría simular todos los pasos a seguir internamente en el sistema en el acceso a la base de datos de clientes cuando se quiere obtener dinero del cajero, pero sin que realmente se trate de una base de datos real ni de un cliente del banco.

• Un programa real que se adecue en parte al software que se desea desarrollar. Por ejemplo, se puede disponer de una aplicación relacionada con un “cajero automático”, que al presentarla al cliente, permita al analista identificar las necesidades del cliente y por lo tanto los requisitos del software a construir.

Normalmente, el prototipo sirve como mecanismo para identificar los requisitos del software, y su construcción suele llevar las siguientes etapas:

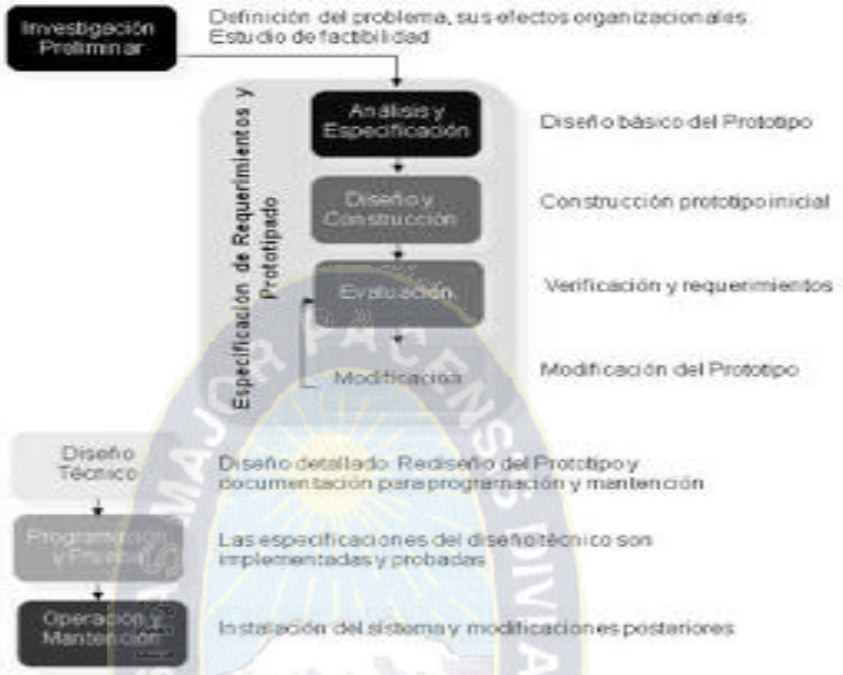


Figura3.18: Etapas del Modelo prototipo

Fuente :Elaboración propia

1) Recolección de requisitos. El ingeniero de software y el cliente definen los objetivos globales del software, y aquéllos más específicos que se desean destacar con el prototipo.

2) Diseño rápido. Centrado en los aspectos del software visible al usuario (por ejemplo, interfaz de usuario, entradas y salidas…).

3) Construcción del prototipo.

4) Evaluación del prototipo. Se realiza por el cliente y usuarios, lo que permitirá concretar y refinar los requisitos del software a desarrollar.

5) Refinamiento del prototipo. Se produce un proceso iterativo en el que el prototipo es refinado para que satisfaga las necesidades del cliente, al tiempo que facilita al ingeniero de software un mejor conocimiento del sistema.

6) Producto. En la mayoría de los casos este sistema refinado (piloto) hay que desecharlo y hacer uno nuevo. Por ello, el desarrollo de un prototipo se debe

planificar con el acuerdo expreso del cliente. Algunos ingenieros del software abogan por desarrollar rápidamente un prototipo que les permita especificar completamente el sistema y obtener más consistentemente el producto final. Sobre el desarrollo rápido de prototipos, pueden realizarse las siguientes observaciones:

• Un prototipo rápido es básicamente una técnica de análisis que permite completar el conjunto de requisitos funcionales de un sistema software.

• Lo deseable es evolucionar el prototipo hasta obtener el producto final, en lugar de deshacerlo y construir un producto final nuevo. Este deseo es válido si del prototipo se puede obtener dicho producto (lo que no suele ser fácil), y su coste es inferior a su reconstrucción. Incluso, se podría recomendar utilizar aquellas técnicas que permitan evolucionar un prototipo hasta el producto final.

• Cualquier aplicación nueva que el ingeniero de software sospeche que su funcionalidad puede presentar el riesgo de no ser aceptable para el usuario o si la interfaz de usuario es importante para el éxito de la aplicación, es una aplicación fuertemente candidata para que se desarrolle un rápido prototipo.

• En un proyecto de prototipo bien planificado, aproximadamente el 50% del esfuerzo de desarrollo, desde su inicio hasta la aprobación final de su funcionalidad, es la contribución del usuario. Los equipos de prototipo están compuestos típicamente por la mitad de usuarios y la otra mitad de desarrolladores software.

• Es habitual tener que tirar la primera versión de cualquier sistema que se desarrolle por primera vez. Por ello, es aconsejable que la primera demostración de un prototipo rápido sea intencionalmente imperfecta, de forma que sea barato de producir y muy fácil de modificar, para que se pueda garantizar que el sistema final que se suministra se ajuste mejor a los requisitos del usuario.

• El prototipo rápido es una solución que “evita el riesgo” en lugar de una solución de riesgo. Así, el prototipo rápido no introduce nuevos riesgos políticos o económicos al proceso de desarrollo de software, sino que reduce

significativamente varios factores de riesgo asociados con su desarrollo, como los que se han señalado anteriormente.

• El prototipo rápido es un método normal para el desarrollo de nuevas aplicaciones y llegará a ser más y más evidente que el prototipo rápido produce mejores sistemas y con costes más bajos.

3.14.1 Ventajas y desventajas del Modelo de" prototipos"

Ventajas:

• Permite la construcción del sistema con requisitos poco claros o cambiantes

• El cliente recibe una versión del sistema en muy poco tiempo, por lo que lo puede evaluar, probar e, incluso, empezar a utilizarlo

• Se pueden introducir cambios en las funcionalidades del sistema en cualquier momento

• Involucra al usuario en la evaluación de la interfaz de usuario

• Se reduce el riesgo y la incertidumbre sobre el desarrollo

• Genera signos visibles de progreso, que se utilizan cuando existe una demanda en la velocidad del desarrollo

• Permite entender bien el problema antes de la implementación final

Desventajas:

• El cliente puede quedar convencido con las primeras versiones y, quizás, no vea la necesidad de completar el sistema o rediseñarlo con la calidad necesaria

• Requiere trabajo del cliente para evaluar los distintos prototipos y traducirlo en nuevos requisitos

• Requiere un tiempo adicional para definir adecuadamente el sistema

• No se sabe exactamente cuánto será el tiempo de desarrollo ni cuantos prototipos se tienen que desarrollar

• Si un prototipo fracasa, el coste del proyecto puede resultar muy caro.

4. EVALUACIÓN DE RESULTADOS

4.1 ANALISIS DE DATOS Y RESULTADOS

La traducción automática (TA) es traducción automatizada. Es el proceso mediante el cual se utiliza software de computadora para traducir un texto de un lenguaje natural (como el chino) a otro (como el español).

Al procesar cualquier traducción, humana o automática, el significado del texto en el idioma original (origen) se debe restaurar totalmente en el de destino, es decir, en la traducción. Aunque en apariencia parezca sencillo, es mucho más complejo.

La traducción no es una mera sustitución de una palabra por otra. Un traductor debe interpretar y analizar todos los elementos del texto y saber cómo influyen unas palabras en otras. Para ello se necesitan amplios conocimientos de gramática, sintaxis (estructura de las oraciones), semántica (significados), etc., de los idiomas de origen y de destino, además de familiaridad con cada región específica.

Tanto la traducción humana como la automática tienen sus propios desafíos. Por ejemplo, dos traductores individuales no pueden producir traducciones idénticas del mismo texto en el mismo par de idiomas, y es posible que se requieran varias rondas de revisiones para lograr la satisfacción del usuario. Pero el mayor desafío reside en cómo se pueden producir traducciones de

calidad aptas para ser publicadas mediante la traducción automática. Este trabajo de investigación fue desarrollado en base a la gramática básica del chino, para su posterior traducción al idioma español se basa en reglas de traducción del idioma chino y español.

La verificación y validación de una cadena de texto, se hizo en base a los autómatas finitos deterministicos. Y para el proceso de traducción se hizo uso del autómata finito traductor modelado a nuestro caso de uso. La verificación de la hipótesis se comprobara en base a las mediciones obtenidas de la operatividad del prototipo y su análisis.

4.1 PRUEBAS

Recordemos que el presente trabajo trata de la traducción del idioma chino al español el cual tiene como hipótesis:

“La aplicación de un modelo de autómata que permita la traducción del idioma chino al español haciendo uso del pīnyīn.”

Para la conclusión de la presente investigación es necesario demostrar esa hipótesis; y procederemos para el mismo aplicar el método de la hipótesis nula a través de la distribución T- estudent.

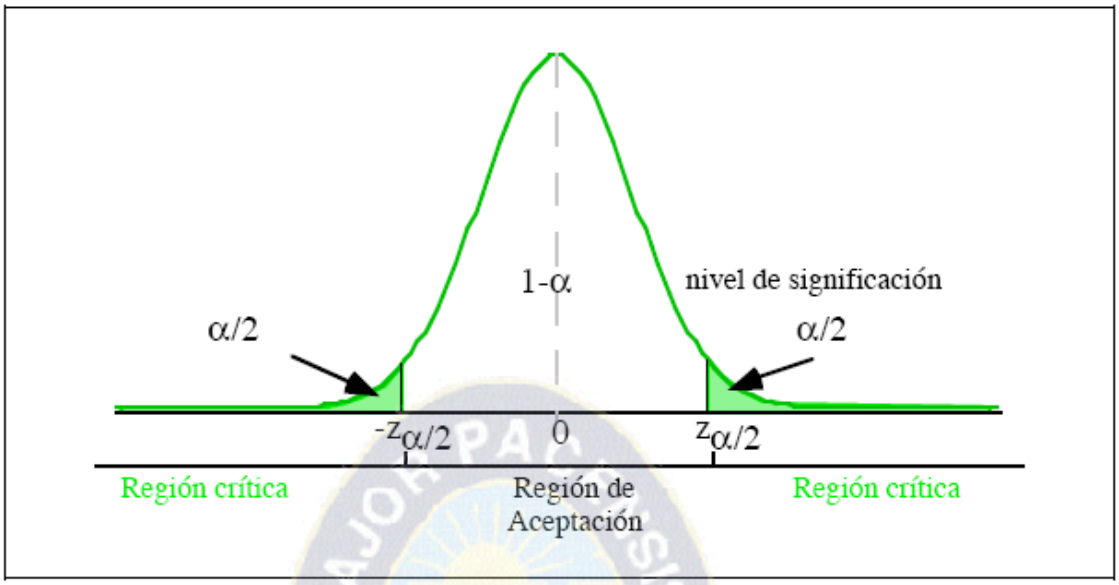


Figura 4.1

Fuente [Elaboración propia]

La prueba t-student o Test-T es una prueba estadística que es utilizada para muestras pequeñas. Y el planteamiento de hipótesis bilateral en la que el estadístico utilizado tiene una distribución t-student si la hipótesis nula es cierta. Se aplica cuando la población estudiada sigue una distribución normal pero el tamaño muestral es demasiado pequeño como para que el estadístico en el que está basada la inferencia este normalmente distribuido, utilizándose una estimación de la desviación típica en lugar del valor real como puede observarse en el grafico anterior.

De donde se calcula de la siguiente manera:

$$Z = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$

μ : porcentaje de la población de estudio

X: porcentaje de la muestra

S: desviación estándar

Considerando a hora como

Ho: “La aplicación de un modelo de autómata que hace uso del pīnyīn correspondiente de los logogramas chinos, permitirá la traducción del idioma chino”

4.2 DETERMINACIÓN DE LA POBLACIÓN

La población que se a estudiado fue proporcionada aleatoriamente de las palabras de curso de chino básico que se encuentran definidas en el lenguaje chino.

4.3 DETERMINACIÓN DEL TAMAÑO DE MUESTRA

Basándose en la teoría de probabilidades y de acuerdo a la población que se tiene, la muestra la determinamos de la siguiente manera:

Para poder evaluar el prototipo realizado y sus resultados se tomara una muestra de 10 palabras y oraciones de cuatro palabras.

No es necesario crear muchas muestras por tamaño, 10 muestras son suficientes.

Clasificación de reconocimiento:

A) buena: los caracteres se encuentran totalmente formados, su ángulo de inclinación es (+/- 12 grados)

B) parcialmente buena: los caracteres pueden presentar pequeñas rupturas, su angulo de inclinación es(+/- 15 grados)

C) mala : los caracteres pueden presentar grandes cortes su ángulo de es mayor a los 15 grados.

A continuación se muestran los principales entrenamientos que se realizan

- Distinción entre caracteres alfabéticos y numéricos:

La letra “B” era reconocida como el número “8” el “0” como la letra “O”,”Q” o ”D”, “U” el “6” como la “G”, el “1” com9 la la “I” entre otras. Como solución se opto por crear un archivo de variables de configuración para especificar lo que se desea reconocer, sea números o letras

Proceso

Se tomaran las oraciones para que puedan ser traducidas como se muestra en el cuadro No. 1

Calculo de la media o promedio
Determinamos la media de la distribución, su formula está dada por

Proceso de calcular la derivación estándar

No. De prueba	No. De palabras	Estado de traducción	X-μ	(X-μ)^2
1	8	8	-0,2	0,04
2	7	5	-3,2	10,24
3	10	9	0,8	0,64
4	8	8	-0,2	0,04
5	6	6	-2,2	4,84
6	9	8	-0,2	0,04
7	7	7	-1,2	1,44
8	10	9	0,8	0,64
9	8	8	-0,2	0,04
10	9	9	0,8	0,64
TOTAL	82	77	-5	18,6

μ =	8,2
X' =	7,7
s=	4,312771731
t=	-3,6661779

Cuadro 1. Evaluación de la hipótesis

Para saber si el valor t (calculado) es significativo, se aplica la formula y se calculo los grados de libertar de la siguiente manera, una vez tenemos cálculo el valor t y los grado de libertad, elegimos el nivel de confianza de 0.01 y se

compara el valor obtenido contra el valor que le correspondería en la tabla t-student

Es decir, en la tabla de la distribución t-student, buscamos el valor t_{n-1} correspondiente con grados de libertad n-1=9 siendo este 3,25.

Por lo tanto, el valor calculado de |t|=3.66 que resulta ser superior al valor de la tabla con un nivel de confianza de 0.01 se llega a concluir la aceptación de la hipótesis de investigación y rechazamos la hipótesis nula.

5. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

Las nuevas tecnologías de la información y la comunicación abren nuevas situaciones de uso lingüístico, dan lugar a nuevas aplicaciones basadas en la lengua y han desarrollado nuevas plataformas de diseminación de la información. Están creando un nuevo escenario, mucho más complejo desde el punto de vista lingüístico, respecto del cual no tenemos todavía la suficiente distancia ni perspectiva como para aventurar cuáles serán sus efectos futuros.

La comprensión del significado de las palabras no tiene relación alguna con la traducción. Lo importante es comprender el significado de un texto en un idioma de origen, para obtener un texto con significado equivalente en el otro idioma; es decir, consiste en transferir el significado del lenguaje de una lengua a otra y el sentido (intención) pretendido.

Los sistemas de traducción basados en el conocimiento; requieren del desarrollo de gramáticas y léxicos monolingües y bilingües, así como de reglas de transferencia que permiten establecer equivalencias entre las estructuras de las lenguas que se tratan.

La diferencia es que el ordenador trabaja de modo sintáctico mientras que nuestros cerebros lo hacen además de modo semántico. Con esto quiere decir que la mente humana dota de significado a los símbolos con los que trabaja

mientras que los ordenadores no lo hacen. Para el ordenador las frases introducidas (que pretenden medir su inteligencia) no son más que secuencias de unos y ceros sin el menor significado. De hecho, el programa que el ordenador sigue para obtener una respuesta maneja esos símbolos como secuencias sin sentido, y no hace referencia al significado de dichas frases.

Es por eso que no se puede obtener un 100% de fiabilidad respecto a un traductor automático.

Por otro lado fue muy difícil conseguir software de desarrollo que permitiera trabajar con el idioma chino y si corría, no llegaba a funcionar en algunas plataformas que tenían soporte limitado sobre dicho idioma.

5.2 RECOMENDACIONES

Investigar el algoritmo en otras plataformas, se puede aumentar la eficiencia computacional acortando los tiempos de las aplicaciones OCR.

Investigar a mas profundidad sobre la gramática de la lengua china.

La brecha digital que existe entre los dos idiomas a veces impidió poder proporcionar resultados a tiempo; por eso es necesario contar en la biblioteca de informática de software legal y completo para las investigaciones que se hacen con respecto a otros idiomas.

BIBLIOGRAFIA

[HIL09] José R. Hilera González, Juan P. Romero Villaverde, José A. Gutiérrez de Mesa. SISTEMA DE RECONOCIMIENTO ÓPTICO DE CARACTERES (OCR) CON REDES NEURONALES,2009. http://www.cc.uah.es/hilera/docs/1996/c_jiacse1/c_jiacse1. Htm.

[CAN03] Javier Cano y Juan Carlos Pérez.OCR (Optical Character Recognition) http://zweb.iti.upv.es/services/reviewtic/public/2003/11/2003-11-ocr.

[ROD04] José Ramón Rodón Ortiz, Javier Ráez Rus, Ismael Vargas Pina. 2004.Trabajo dirigido:“OCR basado en árboles binarios”. OCR para caracteres impresos basados en árboles binarios. http://alojamientos.us.es/gtocoma/pid/pid10/OCRarbolbinario.html

[ALF07] Enrique Alfonseca Cubero, Manuel Alfonseca Cubero, Roberto Moriyón Salomón." Teoría de autómatas y lenguajes formales". McGraw-Hill (2007). Capítulos 3 y 7.

[HOP05] John E. Hopcroft, Rajeev Motwani, Jeffrey D.Ullman. "Introducción a la teoría de autómatas, lenguajes y computación" (3ª edición). Ed, Pearson Addison Wesley,2005.
Sects. 2.1-2.2; Sects. 2.3-2.8; HMU, Chap. 4;HMU, Sects. 3-1-3.

[ALF97] Manuel Alfonseca, Justo Sancho, Miguel Martínez Orga. "Teoría de lenguajes, gramáticas y autómatas". Publicaciones R.A.E.C. 1997 Capítulos 4,5,y 8

[BEI10] Beijin Language and cultura University Press.”El nuevo libro de chino practico” (3ra Edición) 2010.

[YIN09] Yip Po-Ching y Don Rimmington.”Gramatica básica del chino” teoría y práctica. 2da Edición publicada en el 2009.

[SUN00] Sun yizhen, LuChuanshan, Ma Mingwei,Wang Xiaowen Yang Zhiying, Chen Quan, Lian Meijin “Nuevo diccionario Chino -Español” (3ra Edición) Editorial XĪN HÀN XĪ CÍDIĂ, 2000.

[CIR06] Universidad de Granada, Dr. Juan Jóse Ciruela Alférez.”El chino de Hoy” Ed. Univesity Granada Press, 2006.

[ASS11] ASSIMIL para hispano hablantes.”El chino de bolsillo” 2da edición 2011.