

Benemérita Universidad Autónoma de Puebla  
Facultad de Ciencias de la Computación



**“TRADUCCIÓN AUTOMÁTICA DE TEXTO A  
PICTOGRAMAS”**

---

Tesis para obtener el grado de  
**Maestro en Ciencias de la Computación**

presenta:

**Luis Gustavo Olivares Soto**

Asesor de Tesis:

**Dr. David Eduardo Pinto Avendaño**

Co-asesor de Tesis:

**Dra. Maria Josefa Somodevilla Garcia**

Puebla  
Junio 2019

# Dedicatoria

*Esta tesis es un logro mas que llevo a cabo, dedico esta tesis a mi padre y hermanos  
pero especialmente a mi madre por estar conmigo en todo momento y apoyarme  
cuando mas lo necesitaba,  
Gracias por todo familia los quiero mucho*

# Agradecimientos

Agradezo al *Consejo Nacional de Ciencia y Tecnología* (CONACYT) por la beca otorgada a lo largo de la elaboración de la presente, del proyecto titulado “Traducción Automática de Texto a Pictograma”.

Un gran agradecimiento a mi asesor de tesis *Dr. David Eduardo Pinto Avendaño* por todo el tiempo dedicado en este proyecto así como su paciencia, entrega y motivación para que yo sea una mejor persona y estudiante. A mi co-asesora la Dra. Maria Somodevilla García por la enseñanza en todo el proceso de la maestría que gracias a ella aprendi muchas cosas.

Gracias a mi padre por siempre apoyarme y motivarme a seguir adelante a mis hermanos por algunas veces ser duros con migo pero gracias a ellos eh llagado hasta donde estoy, y principalmente gracias a mi madre por su apoyo incondicional y a enseñarme que puedo hacer las cosas de la mejor manera para salir adelante.

Por último un agradecimiento agradezco a mis amigos por apoyarme en especial a Karen que me ha ayudado en el desarrollo la profesional y sobretodo personal en este tiempo.

*Luis Gustavo Olivares Soto ¡GRACIAS!*

# Índice general

Dedicatoria	I
Agradecimientos	I
Lista de figuras	V
Lista de tablas	VI
Resumen	VII
<b>1. Introducción</b>	<b>1</b>
1.1. Planteamiento . . . . .	1
1.2. Preguntas de investigación . . . . .	2
1.2.1. Objetivos . . . . .	2
1.3. Hipótesis . . . . .	3
1.4. Impacto . . . . .	3
1.5. Resultados esperados . . . . .	3
1.6. Estructura de la Tesis . . . . .	4
<b>2. Estado del arte</b>	<b>5</b>
<b>3. Marco teórico</b>	<b>10</b>
3.1. Procesamiento de Lenguaje Natural . . . . .	10
3.2. Pictogramas . . . . .	11
3.3. Ambigüedad . . . . .	11
3.4. Tablas hash . . . . .	12

---

3.5. Análisis de texto . . . . .	13
3.6. Precisión . . . . .	13
<b>4. Metodología de solución</b>	<b>15</b>
4.0.1. Texto como pictograma . . . . .	15
4.0.2. Traducción automática de texto a pictogramas . . . . .	19
4.0.3. Ambigüedad . . . . .	20
<b>5. Resultados experimentales</b>	<b>24</b>
5.1. Conjunto de datos . . . . .	24
5.2. Evaluación . . . . .	25
<b>6. Conclusiones</b>	<b>27</b>
<b>Bibliografía</b>	<b>28</b>

# Índice de figuras

4.1. Ejemplo de tabla hash para la identificación de pictogramas. . . . .	16
4.2. Representación gráfica del etiquetado de pictogramas. . . . .	18
4.3. Traducción de texto a pictogramas presentado gráficamente. . . . .	20
4.4. Ejemplo gráfico de ambigüedad en las palabras. . . . .	21
4.5. Funcionamiento de aplicación web. . . . .	23

# Índice de tablas

4.1. Etiquetado <i>POS</i> . . . . .	17
4.2. Palabras ambiguas contenidas en el corpus de pictogramas . . . . .	19
5.1. Clasificación del total de pictogramas en el corpus. . . . .	25
5.2. Clasificación de pictogramas sin repetir, localizados en el corpus. . . .	25
5.3. Porcentaje de efectividad en la traducción texto a pictogramas . . . .	26

# Resumen

En esta tesis, estudiamos la traducción de texto a pictogramas que son imágenes o signos que representan de manera más o menos realista un cierto significado. Su objetivo es traspasar la expresión lingüística para expresar figurativamente un objeto o un significado. Dado que los pictogramas son construidos para ser comprendidos en a lo más tres miradas evitándose detalles superfluos, los mismos pueden ser usados como una forma primitiva de comunicación entre personas que tengan dificultades para entender o manejar un lenguaje más amplio o complejo como el lenguaje natural.

También se abordó el problema de traducir texto a pictogramas, con la finalidad de generar una plataforma mediante la cual una persona normal pueda comunicar con personas que tienen problemas de desarrollo intelectual en su proceso de comunicación. En este sentido, se estudiaron modelos básicos o primitivos de comunicación para determinar el tipo de componentes lingüísticos primitivos que resultaron ser viables en el proceso de comunicación. Se analizaron los vínculos entre los componentes morfológicos verbo, sustantivo y adjetivo con respecto a un corpus de pictogramas estándar, con la finalidad de determinar su uso práctico en ambientes reales y de acuerdo con diferentes niveles de discapacidad en el desarrollo intelectual.

# Capítulo 1

## Introducción

Existen muchos medios de apoyo a la comunicación entre personas pero no mediante pictogramas los cuales son imágenes que representan una frase, acción sentimientos, etc.

En este trabajo se presenta un traductor automático de texto a pictogramas para ayudar a personas que presentan alguna discapacidad en su proceso de comunicación con otra personas.

Al paso de los años, se realizan investigaciones sobre la importancia o impacto que tienen los pictogramas en personas con dificultades en su proceso de la comunicación, sin embargo a la fecha no se han desarrollado sistemas con la utilización de pictogramas que sean capaces de ayudar a estas personas.

En este trabajo, se utilizan técnicas recuperación de información y procesamiento de lenguaje natural para poder obtener resultados precisos al momento de la traducción a pictogramas.

### 1.1. Planteamiento

El problema que dio origen a este proyecto es el de una falta de apoyo a personas con problemas de discapacidad en su desarrollo intelectual y que claramente podrían beneficiarse de un sistema de traducción automático entre texto y pictogramas, puesto que las personas normales podría expresarse de manera normal (usando lenguaje natural), y dicho sistema traduciría el mensaje a un conjunto de pictogramas que

serían entendidos de manera mucho más fácil que el mensaje original expresado en lenguaje natural. Existen en la actualidad personas que presentan dificultades para comunicarse ya sea de manera oral y/o escrita y para las cuales sería de mucho beneficio el utilizar los pictogramas (imágenes) con la finalidad de expresar ideas, conceptos, sentimientos, etc. Habilitaría a la persona para poseer una manera de interpretar, comprender y transformar su realidad en imágenes, para qué, a través de éstas exprese y transmita su pensamiento al interlocutor.

## 1.2. Preguntas de investigación

En este apartado se en listan las preguntas que dieron origen a este trabajo de tesis.

¿Como podrían establecer un puente de comunicación una persona con otra que presenta discapacidad intelectual al momento de comunicarse?

¿Son entendibles los pictogramas para las personas que se comunican normalmente?

¿Son los pictogramas una alternativa eficiente para comunicarse?

¿Que efecto tienen los pictogramas al momento de querer expresar ideas, sentimientos, conceptos, etc. a personas con problemas de comunicación?

¿Ayudaría la traducción de texto a pictogramas a personas que tienen problemas para comunicarse?

¿Se puede traducir el texto a pictogramas sin presentar ambigüedad?

### 1.2.1. Objetivos

Los objetivos se plantearon en objetivo general y objetivos específicos, se describen a continuación.

#### Objetivo general

- Diseñar y construir un modelo computacional para la traducción automática de texto a pictogramas.

### **Objetivos específicos**

- Determinar el tipo de técnicas necesarias para la construcción de un modelo computacional para la traducción automática de texto a pictogramas
- Seleccionar y evaluar un corpus de pictogramas que sea útil en el proceso de construcción del modelo de traducción propuesto
- Analizar los vínculos entre los componentes morfológicos verbo, sustantivo y adjetivo con respecto a un corpus de pictogramas estándar, con la finalidad de determinar su uso práctico en ambientes reales y de acuerdo con diferentes niveles de discapacidad en el desarrollo intelectual.
- Construir un modelo de traducción automática de texto a pictogramas
- Diseñar un conjunto de pruebas que evalúen el rendimiento del modelo de traducción construido

### **1.3. Hipótesis**

A continuación se describe la hipótesis dados los cuestionamientos anteriores.

Es posible traducir un texto de dominio de comunicación simple con una tasa de encuentro superior al 40 % de efectividad, con el fin de ayudar a personas establecer un dialogo con otra que presenta alguna discapacidad intelectual en su proceso de comunicación.

### **1.4. Impacto**

El principal aporte es establecer un puente de dialogo entre personas que no presentan alguna limitación pero desean comunicarse con aquellas que manifiestan discapacidades que afectan su comunicación.

### **1.5. Resultados esperados**

- Permitir el máximo porcentaje de efectividad al momento de traducir un texto.

- Desarrollar traducido a partir de un texto a pictogramas evitando ambigüedades.
- Implementar un sistema con la interfaz y funcionamiento adecuado.

## 1.6. Estructura de la Tesis

Este documento se organiza de la siguiente forma:

- Capitulo 1 Se presenta la introducción así como el planteamiento del problema, la hipótesis y los resultados esperados.
- Capitulo 2 Se exponen los trabajos desarrollados en el área de "Traducción automática de texto".
- Capitulo 3 Marco Teórico se describe la definición de los conceptos utilizados en esta tesis.
- Capitulo 4 En esta sección se muestra la solución al problema planteado.
- Capitulo 5 El conjunto de datos y los resultados obtenidos se encuentran en este capitulo.

Finalmente en el Capitulo 6 se ofrecen las conclusiones de este trabajo despues de este capitulo se listan las referencias utilizadas.

# Capítulo 2

## Estado del arte

En los últimos años se ha analizado la forma de ayudar a personas con problemas de desarrollo intelectual y con discapacidades en lectura y/o escritura. Las investigaciones muestran que usar pictogramas en lugar de texto es una herramienta valiosa para ayudar a estas personas. Por esta razón, en esta sección enumeramos algunos trabajos reportados en la literatura que están vinculados con el tema de tesis propuesto y que tiene que ver con la construcción de un modelo de traducción automático de texto a pictogramas.

En [18] se explica que la traducción simplificada de texto a pictograma ayuda a personas con discapacidades intelectuales con el fin de habilitarles o facilitarles la comunicación con otras personas. El sistema de traducción de texto a pictografía presentado en este trabajo de investigación traduce automáticamente el texto escrito holandés en una serie de pictografías Sclera o Beta. El sistema de referencia presenta al lector una traducción de pictograma por palabra más o menos literal. Como resultado, las oraciones de entrada largas y complejas conducen a largas y complejas traducciones de pictogramas, dejando a los usuarios confundidos y distraídos. Para superar estos problemas, los autores desarrollaron un sistema de simplificación basado en reglas para la traducción holandesa de texto a pictograma. Las evaluaciones que los autores presentan muestran una gran mejora sobre los baselines.

En [23] se presenta un procedimiento para convertir una cadena entrante que es e encuentra fonética codificada y que está representada en un idioma de Asia oriental, en sus caracteres de texto apropiados. La conversión se desencadena cuando el

análisis dinámico de IME de la cadena indica un umbral de probabilidad de conversión correcta. La conversión de texto fonético no convertido también se desencadena cuando el usuario reubica el punto de inserción. El usuario puede seleccionar porciones de texto afectadas por el error de conversión y acceder a una cadena fonética editable correspondiente y una lista priorizada de conversiones alternativas para esa cadena. La cadena fonética correspondiente se genera a partir del texto seleccionado si la cadena de entrada original no está disponible.

También en [10] se explica que una imagen puede ser una poderosa herramienta de comunicación. Sin embargo, crear imágenes para ilustrar las instrucciones del paciente puede ser una tarea costosa y que requiere mucho tiempo. Sobre la base de la investigación previa en esta área, los autores desarrollan una aplicación informática que convierte automáticamente el texto en imágenes utilizando el procesamiento del lenguaje natural y las técnicas de gráficos por computadora. Después de las pruebas iterativas, el sistema de ilustración automatizado se evaluó utilizando 49 instrucciones de descarga de cardiología que no se habían visto anteriormente. Tres evaluadores evaluaron la integridad de las ilustraciones generadas por el sistema utilizando una escala de tres niveles. El acuerdo promedio entre evaluadores para el texto correctamente representado en el pictograma fue de alrededor del 66 por ciento. Como la ilustración en este contexto está destinada a mejorar el texto en lugar de reemplazarlo, estos resultados respaldan la posibilidad de realizar una ilustración automatizada.

En [5] se explica que una imagen puede ser considerada en traducción como un paratexto icónico cuya lectura e interpretación simbólicas requieren de una cierta disciplina metodológica traductiva a la hora de cultivar el sentido de la mirada del traductor. A través de dos corpus diferentes, por un lado, las imágenes dibujadas de unos supuestos pictogramas de la Xunta de Galicia y, por otro, las nuevas caras de los estuches de Kinder Chocolate, José Yuste Frías demuestra en este artículo que la imagen en traducción no es universal. La manipulación de la imagen conlleva no sólo implicaciones simbólicas, sino también, ideológicas, políticas, sociales y culturales que influyen enormemente en la presentación final de los productos de la traducción.

En [26] se dice que la inclusión social de personas con discapacidades intelectuales y del desarrollo puede promoverse ofreciéndoles formas de usar Internet de forma

independiente. Las personas con discapacidades de lectura o escritura pueden usar pictografías en lugar de texto. Se presenta también un recurso en el que se vincula un conjunto de 5,710 pictografías con conceptos semánticos léxicos en Cornetto, una base de datos similar a Wordnet para holandés. Se muestra que, al utilizar este sistema de traducción de texto a pictografía de recursos, es posible mejorar en gran medida la comparación de la cobertura con una línea base en la que las palabras se convierten en pictografías solo si la palabra es igual al nombre del archivo.

En [22] presentan un sistema de traducción Pictograph-to-Text para personas con Discapacidades Intelectuales o de Desarrollo (IDD). El sistema traduce los mensajes de pictografía, que consisten de una o más imágenes en un texto en holandés, utilizando enlaces de WordNet y un modelo de lenguaje basado en n-gramas. También se proporcionan varios métodos de información de pictogramas que ayudan a los usuarios a seleccionar los pictogramas apropiados. Si bien, este trabajo lleva a cabo el proceso inverso al requerido, podría ser de utilidad en la construcción de la propuesta.

En [21] describen cómo se adaptó un sistema de traducción de texto a pictograma holandés, diseñado para aumentar el texto escrito para personas con discapacidades intelectuales o del desarrollo (IDD), para poder utilizarlo en inglés y español. El sistema original tiene un diseño independiente del idioma. En lo que se refiere a la parte textual, es adaptable a todos los lenguajes naturales para los cuales están disponibles los enlaces interlinguales de WordNet, los lematizadores y los etiquetadores de parte de la oración. En lo que se refiere a la parte pictográfica, puede ser modificada para varios lenguajes pictográficos. Las evaluaciones de dicho trabajo de investigación muestran que los resultados están en línea con el rendimiento del sistema holandés original. La traducción de texto a pictograma tiene un amplio potencial de aplicación en el dominio de la Comunicación Aumentativa y Alternativa (AAC).

En el trabajo [27] se deduce que tanto chistes como cómics conforman un potencial material con una información lingüística y cultural extraordinarias para consolidar aspectos léxicos, semánticos y morfo-sintácticos del idioma alemán. Por todo ello, los cómics debieran incluirse dentro de la taxonomía pedagógica con el fin de diversificar el acervo textual -al que muy a menudo recurre el profesor- y crear mayor dinamismo en la clase de alemán.

También en [17] se emplean chistes lingüísticos como instrumento para la enseñanza del español a extranjeros, con la finalidad de que el alumno aprenda correctamente alguna lengua, lo realizan por medio de cómics, ya que son un medio de expresión, el empleo de la tira cómica sirve como material de apoyo en las clases de gramática y conversación, en su trabajo se obtuvieron generalmente buenos resultados.

De igual manera en [16] realizan la estimulación lingüística a través de gráficos e imágenes. Son aplicados a la didáctica de la lengua en la Educación Secundaria, ya que los gráficos en el uso por el alumno son un medio idiosincrasia, que revela su estructura de conocimientos, incluso sus actitudes, creatividad, etc. Señalan la forma diferente de hacer diagramas de estudio o de trabajo con subgrafismos que tienen los alumnos con el cual es posible practicar formas alternativas de evaluación conjunta, auto evaluación, etc.

En otro artículo [12] señalan que utilizan los pictogramas en las actividades de lectura y escritura, aunque de menor modo en actividades de rutina y de juego, pues deben utilizarlas en todas las áreas sobre todo en las del juego ya que los niños no se sienten aburridos al aprender y demuestran que existe un pequeño número, pero significativo de niños y niñas que no describen con facilidad ciertas características de una lámina con objetos y gráficos.

El trabajo presentado en [20] utilizan los pictogramas en el proceso educativo a través de las actividades pedagógicas planificadas para las clases diarias, facilitan así la estructuración del lenguaje según el desarrollo de los niños y niñas, y su correspondiente evaluación por el impacto causado en la comunidad educativa. La utilidad de los pictogramas en los procesos de lectura, escritura y otras formas del lenguaje la utilizan como una secuencia de actividades que se realizarán durante el día.

En el artículo [9] los pictogramas son de gran impacto dentro del desarrollo de la Expresión Oral y Escrita, puesto que ayuda a que los niños de 4 a 5 años desarrollen una gran capacidad y variedad de su lenguaje aumentando su facilidad de expresarse y desenvolverse de mejor manera. Su utilidad se refleja en que los Pictogramas aumenta la atención de los niños, activa la memoria ya que ellos pueden anticipar imágenes que van apareciendo ayudando a trabajar conceptos como nociones temporales y espaciales asociando lugares con personas, actividades, etc. Siendo útil dentro

de la Expresión Oral y Escrita ayudando a desarrollarlo.

En [4] muestran un trabajo crítico-visual que refleja el estatus contemporáneo de los emoticones y los emojis, signos de una creciente integración de lo digital en la vida cotidiana. Entendido a la luz de la producción reciente del artista, el lenguaje analítico visual de Pere Joan funciona como una aceptación de las posibilidades formales señaladas por lo digital y a la vez como una reivindicación de la cultura analógica y la reflexión histórica.

# Capítulo 3

## Marco teórico

### 3.1. Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (*PLN*) es un área o campo de la inteligencia artificial que combina las tecnologías de la ciencia computacional con la lingüística aplicada.

Segun [15] el procesamiento de lenguaje natural es una subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional que tiene como razón principal construir sistemas y mecanismos que permitan la comunicación entre personas y maquinas por medio de lenguajes naturales; el *PLN* busca crear programas que puedan analizar, entender y generar lenguajes que los humanos utilizan habitualmente, de manera que el usuario pueda llegar a comunicarse con la maquina o computadora de la misma forma que lo haría con un ser humano. El objetivo principal de este campo es conseguir que las computadoras realicen tareas que impliquen el lenguaje humano, tareas como comunicacion humano-computadora, mejorando la comunicacion entre personas, o simplemente hacer uso del procesamiento de texto en el lenguaje [14].

El *PLN* consiste en la utilizacion de un lenguaje natural para comunicarnos con la computadora, la cual debe entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje [28].

Se entiende, entonces que *PLN* es la habilidad que tiene la maquina para procesar la informacion transmitida, no simplemente de las letras o los sonidos del lenguaje, es decir, en este sentido por ejemplo un perico no es un animal parlante; asi, una contestadora telefonica comun, una impresora, un procesador de palabras tampoco son dispositivos o *software* de *PLN*, mientras que un traductor automatico sin duda lo es [6].

El *PLN* es la parte de la Inteligencia Artificial (I.A.) que se define como: “La ciencia y la ingenieria de hacer maquinas inteligentes” [2].

## 3.2. Pictogramas

Las Tecnologías de la Información y las Comunicaciones (TICs) y la WWW deben ser una fuente de oportunidades para la integración, el aprendizaje, el empleo y no un conjunto de nuevas barreras que aumente la exclusión y la discriminación.

Un pictograma es una herramienta que permite la representación del lenguaje mediante dibujos, fotos e imágenes. Los símbolos gráficos abarcan desde sistemas muy sencillos basados en dibujos o fotografías hasta sistemas progresivamente más complejos como los sistemas pictográficos. Tambien resultan muy útiles también para las personas que no están alfabetizadas a causa de la edad o la discapacidad. Debe considerarse que en muchos casos las personas con discapacidades motoras y/o cognitivas, disponen de profesionales o acompañantes terapéuticos que los asisten y con los cuáles generalmente establecen y organizan estos sistemas de símbolos gráficos.[8].

Estos sistemas son aptos no sólo para personas con discapacidad intelectual o Trastornos del Espectro Autista (TEA), sino también por personas con discapacidades motoras como Parálisis Cerebral (PC), Esclerosis Lateral Amiotrófica (ELA), Edad Mental (EM)[7].

## 3.3. Ambigüedad

La ambigüedad se refiere al estado de tener o expresar más de un significado posible o algo abierto a más de un significado posible. Se refiere al estado en el que

una palabra o una declaración, cualquier entidad lingüística, se puede entender de más de una manera. Es el estado donde es difícil localizar un significado preciso o proporcionar una explicación, ya que implica muchos significados diferentes. La ambigüedad es la falta de claridad en virtud de tener más de un significado. Con respecto a la interpretación, la ambigüedad es la duda o la incertidumbre. Se refiere a algo de significado dudoso. Una palabra, una frase, una oración u otros objetos lingüísticos se llaman ambiguos si pueden interpretarse razonablemente de más de una manera. El caso más simple es una sola palabra con más de un sentido. La palabra "banco", por ejemplo, puede significar "institución financiera", "borde de un río" u otras cosas [13].

Para una solución a la ambigüedad el etiquetado automático de texto es un primer paso importante en descubriendo la estructura lingüística de grandes corpus de texto. La información de parte del discurso facilita el análisis de alto nivel, como reconocer frases nominales y otros patrones en texto. Una forma de modelo de Markov ha sido ampliamente utilizada que asume que una palabra depende probabilísticamente solo de su categoría de parte del discurso, que a su vez depende únicamente de las categorías de las dos palabras anteriores, es decir, estimación de parámetros. Se ha utilizado con este modelo [1].

### 3.4. Tablas hash

Un método utiliza una pluralidad de tablas hash para proporcionar un repositorio de objetos para el desarrollo y uso de aplicaciones orientadas a objetos. El método incluye almacenar un identificador de objeto y una representación del objeto en una primera tabla hash y almacenar datos sobre el objeto y el identificador de objeto en una pluralidad de tablas hash emparejadas con las tablas hash organizadas en pares de tablas duplicadas. Los datos sobre el objeto incluyen el nombre de clase de un objeto, los métodos de objeto, los tipos de retorno y los valores de datos devueltos por los métodos de objeto. Por lo tanto, las aplicaciones orientadas a objetos se pueden desarrollar sin tener en cuenta si el sistema del usuario final incluye un repositorio de objetos[11].

### 3.5. Análisis de texto

Se suele llamar análisis de contenido al conjunto de procedimientos interpretativos de productos comunicativos ( mensajes, textos o discursos ) que proceden de procesos singulares de comunicación previamente registrados, y que, basados en técnicas de medida, a veces cuantitativas (estadísticas basadas en el recuento de unidades), a veces cualitativas (lógicas basadas en la combinación de categorías) tienen por objeto elaborar y procesar datos relevantes sobre las condiciones mismas en que se han producido aquellos textos, o sobre las condiciones que puedan darse para su empleo posterior. El análisis de contenido, de hecho, se convirtió a finales del siglo XX en una de las técnicas de uso más frecuente en muchas ciencias sociales, adquiriendo una relevancia desconocida en el pasado a medida que se introdujeron procedimientos informáticos en el tratamiento de los datos.[19].

El propósito fundamental de tal gramática del texto era el de intentar proporcionar una descripción explícita de las estructuras (gramaticales) de los textos. La tarea más obvia de tal descripción era explicar las relaciones (semánticas) de la coherencia entre las oraciones, y otros aspectos fundamentales del discurso. Así como la gramática de la oración necesita hacer explícita la forma en la que las cláusulas de oraciones complejas se relacionan semánticamente[25].

### 3.6. Precisión

Precisión es una medida de evaluación común en la recuperación de información, se basa en la comparación de un resultado esperado y el resultado efectivo del sistema evaluado [3]. Esta medida ha sido adaptada para la evaluación de porcentaje de efectividad al momento de la traducción de la traducción de texto a pictogramas. *Precisión* mide la relación de pictogramas encontrados correctamente sobre el total de palabras ingresadas, Que se mide como en la formula (3.1).

$$Precision = \frac{pe}{pi} * 100 \quad (3.1)$$

Donde:

$p_e$ : palabras encontradas

$p_i$ : Palabras ingresadas

Dando como resultado el porcentaje de efectividad de traducción de texto a pictogramas.

# Capítulo 4

## Metodología de solución

En esta sección se muestran y describen detalladamente las soluciones y métodos llevadas a cabo para la solución a la traducción de texto a pictograma. Como se mencionó anteriormente la ambigüedad fue el problema principal a resolver en esta investigación, a continuación se muestran las fases realizadas

### 4.0.1. Texto como pictograma

Para la representación de texto como pictogramas se realizó un indexado, descrito a continuación:

1. **Selección de pictogramas** En un principio cada pictograma del corpus utilizado tenía asociado un nombre que describe la imagen, para que un texto sea representado como imagen se considero necesario:
  - (a) Re-estructuración de su nombre. Este paso es uno de los más importantes debido a que el nombre final es la palabra que se traduce, para esto se depuro el nombre original del pictograma, eliminando todos los signos que pudiera tener, como acentos y signos de puntuación.
  - (b) Separación de conceptos y palabras individuales: Un pictograma puede representar una sola palabra (*por ejemplo: casa, carro, libro, etc*) ó una serie de palabras (*por ejemplo: jugar al golf, cama de gato*) se realizó un filtrado de palabras y conceptos dividiéndolo en dos tablas que permitió

lograr una mejor traducción a pictogramas aumentando así su porcentaje de efectividad considerablemente. Se localizaron 2817 conceptos.

- (c) Asignación de índice: En esta fase se realizó una tabla hash que posteriormente es transformada a una tabla en la base de datos creada en mysql. Con una tabla hash se proporcionó una búsqueda de pictogramas más rápida, un ejemplo de la tabla se muestra en la figura 4.1.

Figura 4.1: Ejemplo de tabla hash para la identificación de pictogramas.

Pictograma	ID
enfermero	768
hola	123
como	278
estas	32
casa	2345
correr	456
trabajar	23
hamburguesa	983
hija	789
baño	489

- (d) Definición del pictograma: Al tener un pictograma asociada una palabra o concepto que lo representa, se le asignó una definición a este, por ejemplo: casa Edificio o parte de él para vivir. Se utilizó el diccionario RAE y además se realizó un scrapping o filtrado de las definiciones del recurso online de ARASAAC <sup>1</sup>.
- (e) Etiquetado POS (*PartOfSpeech* por sus siglas en inglés) del pictograma: También es importante etiquetar cada pictograma respecto a la parte de oración que corresponde (sustantivo, verbo, adjetivo, etc) en la tabla 4.1 se presenta las categorías encontradas con su respectiva cantidad de pictogramas en cada categoría. Para obtener el mayor porcentaje posible de efectividad se utilizó el diccionario RAE, ARASAAC, un pos tagging de

<sup>1</sup><http://www.arasaac.org>

Python y además un etiquetador online Linguakit <sup>2</sup> creando así un etiquetador propio para que pudiera tener un mayor alcance. El proceso del etiquetado de pictogramas, así como el del etiquetador propio se muestra en la figura 1e.

Tabla 4.1: Etiquetado *POS*

Categorías	Cantidad de pictogramas por categoría
Adjetivo	1780
Adverbio	254
Artículo	12
Conjunción	36
Determinante	31
Número	226
Preposición	44
Pronombre	137
Sustantivo	8759
Verbo	2456

---

<sup>2</sup><https://linguakit.com>

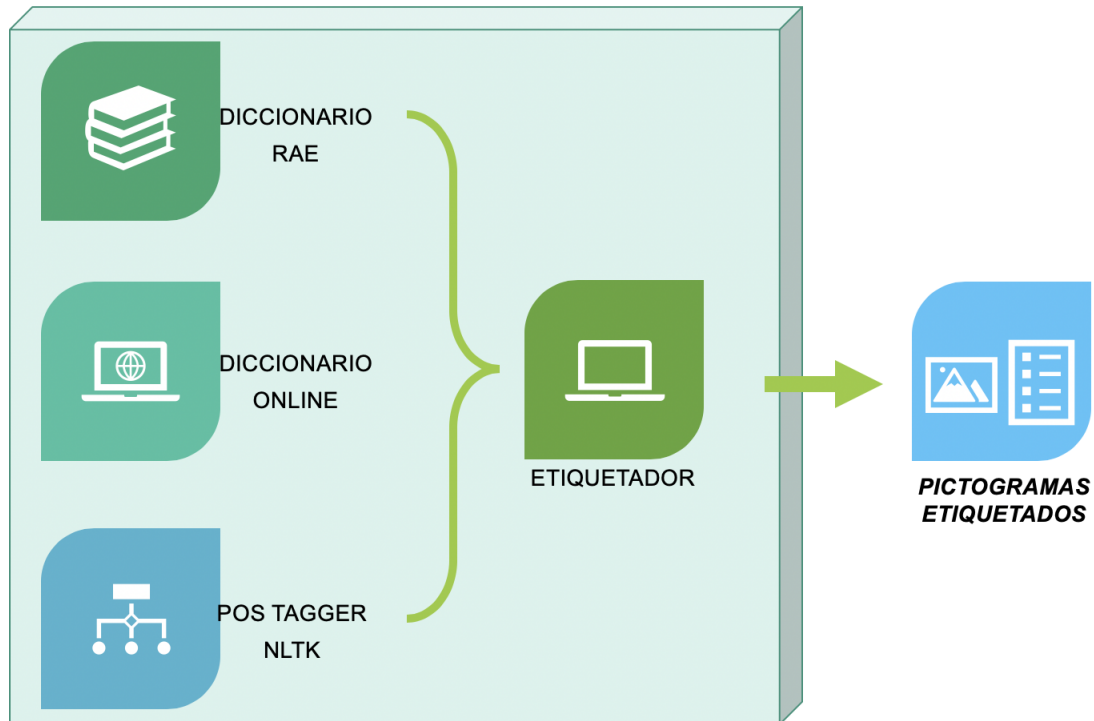


Figura 4.2: Representación gráfica del etiquetado de pictogramas.

- (f) Búsqueda de palabras ambiguas: De todos los pictogramas obtenidos del corpus utilizado, se realizó una búsqueda profunda y detallada de todos aquellos pictogramas que pudieran ser ambiguos, se creó una tabla de igual forma con un índice para cada palabra, así como su etiqueta *POS*, y las palabras que pueden diferenciar las palabras ambiguas. Por ejemplo: 1. banco (referente a un banco de dinero) dinero, cajero, retirar, depositar. 2. banco (referente a un banco de asiento) asiento, sentar, madera. En la tabla 5.2 se enlistan las palabras con ambigüedad y la cantidad de definición para cada una, así como su etiqueta *POS*. La finalidad de este paso, es resolver mejor la ambigüedad la cual será detallada en la siguiente sección.

Tabla 4.2: Palabras ambiguas contenidas en el corpus de pictogramas

Palabra	Cantidad de definición por palabra	POS
Apuntar	2	Verbo
Banco	3	Sustantivo
Carta	1	Sustantivo
Copa	2	Sustantivo
Gato	2	Sustantivo
Lima	2	Sustantivo
Lista	2	Adjetivo, Sustantivo
Pila	2	Sustantivo
Planta	2	Sustantivo
Radio	4	Sustantivo
Ruta	2	Sustantivo

#### 4.0.2. Traducción automática de texto a pictogramas

Obtenido lo detallado anteriormente se realizó el proceso principal de esta tesis, el proceso consta de las siguientes fases:

1. **Pre-procesamiento de texto** El usuario introduce el texto que desea traducir en su lenguaje natural, no importando ortografía de acuerdo a acentos y signos de puntuación etc. Por tal motivo, se toma ese texto y se realiza el pre procesamiento que consta de:
  - (a) Depuración de texto: Se eliminan los acentos y signos de puntuación con el fin limpiar la palabra y mejorar la búsqueda de pictogramas.
2. **Búsqueda de conceptos** Se realiza en dos partes, descritas enseguida:
  - (a) Se toma todo el texto a traducir y la lista de pictogramas que son conceptos, se realiza la búsqueda y si se encuentra alguno se asigna y las palabras restantes por pictogramas se buscan individualmente.
3. **Búsqueda de palabras ambiguas en el texto a traducir** Este proceso se detalla en la siguiente sección para entender mejor.

4. **Selección de pictogramas** Una vez realizadas las búsquedas y encontradas las palabras del texto asociados a los pictogramas se obtienen los id (identificadores) de cada uno y se buscan en la tabla de la base de datos que contiene las imágenes y definiciones. item Traducción final: Se listan los pictogramas asociados a cada palabra o concepto del texto.

Para comprender mejor el algoritmo, en la figura ?? se muestran gráficamente los pasos.

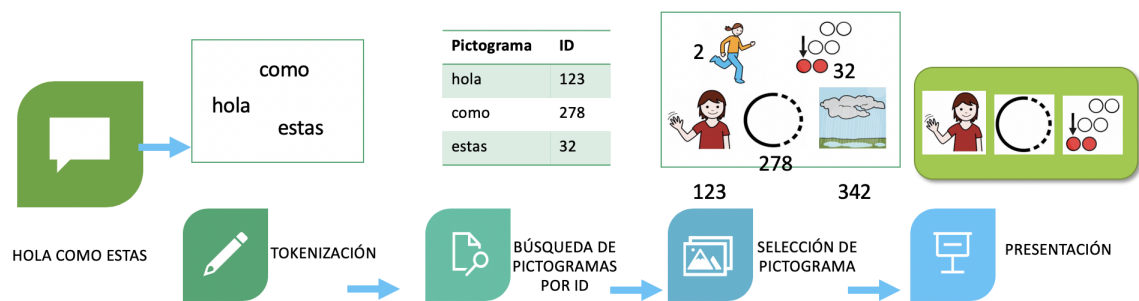


Figura 4.3: Traducción de texto a pictogramas presentado gráficamente.

### 4.0.3. Ambigüedad

El proceso para la solución de ambigüedad en las frases que se traducen, se realizó con lo descrito anteriormente; etiquetado POS y asignación de palabras referentes a la palabra asociada al pictograma, se divide como se describe a continuación.

1. Como su nombre lo indica se accede a la base de datos y se toma la tabla de ambigüedad creada en el proceso anterior y se realiza una búsqueda de todas las palabras del texto a traducir.
2. Si alguna de ellas se ubica en la tabla se procede a volver a tomar el texto y enseguida se evalúa si alguna de las palabras de la oración ingresada se localiza en las palabras asociadas del pictograma y así poner el pictograma acorde a la oración.
3. Para la mejora de la ambigüedad también se consideraron las partes de la oración y se realizó otra búsqueda para averiguar si el texto contiene cada

parte de la oración y una sola de cada una, es decir si el texto contiene un sustantivo, un verbo, un adjetivo, etc.

En la figura 4.4 se muestra un ejemplo de ambigüedad para la palabra "banco" la cual puede tener tres definiciones diferentes.

Pictograma	Diferencia	Definición
	BANCO (peces)	Conjunto de peces
	BANCO (lugar)	Entidad financiera
	BANCO (asiento)	Asiento en el que se pueden sentar

Figura 4.4: Ejemplo gráfico de ambigüedad en las palabras.

### Mejoras de la traducción

Cuando se realiza una traducción, se muestran las estadísticas para conocer la eficiencia del traductor y también para localizar si alguna palabra no pudo ser traducida a pictograma.

1. En el caso de no haber traducido alguna palabra, el traductor muestra al usuario una carta en blanco que permite al usuario agregar en ese momento una imagen o un pictograma que asocie adecuadamente la palabra no traducida para posteriormente tener mejores traducciones..
2. Cuando el resultado es mostrado el usuario puede no entender el significado,

por tal motivo el traductor es capaz de mencionar como audio el significado de cada pictograma con el fin, de poder entenderlo.

### Desarrollo de aplicación web

Como se menciona anteriormente la traducción automática de texto a pictogramas es un procedimiento de gran ayuda para la comunicación con personas que padecen algún tipo de discapacidad del lenguaje, es decir aquellas que no pueden comunicarse de manera habitual, por tal motivo se considero importante llevar esta trabajo a representarlo gráficamente, se realizó una aplicación web programada con html, css y php, brindando interfaces de usuario practicas y de gran utilidad. La forma en la que trabaja la aplicación se muestra gráficamente en la figura 4.5. Además de realizar traducción de texto a pictogramas la aplicación web consta de más utilidades, listadas a continuación.

- Traducción de texto a pictogramas evitando ambigüedad.
- Creación de usuario: Dentro de la aplicación se permite crear usuarios para que cada persona tenga su propio traductor.
- Aumento de pictogramas: A pesar de contar con un gran número de pictogramas, la aplicación permite agregar más, cuando se realiza una traducción y el sistema no localiza un pictograma se permite agregar la imagen en ese momento ó o en cualquier otro desde el apartado agregar.
- Lista de traducciones frecuentes y recientes: También se muestra una lista las traducciones que cada usuario utiliza con más frecuencia ó aquellas que tradujo en un tiempo reciente, permitiendo ahorrar tiempo.
- Traducciones favoritas: Dada la lista de frecuentes y recientes el usuario puede agregar sus traducciones favoritas para almacenarlas y no volver a realizar la traducción desde un inicio
- Voz: Si el usuario no logra entenderlo que significa un pictograma se presenta la opción de escuchar su definición.



Figura 4.5: Funcionamiento de aplicación web.

# Capítulo 5

## Resultados experimentales

En este capítulo se presentan los resultados obtenidos de la tesis en general.

### 5.1. Conjunto de datos

Actualmente existe un gran número de corpus, muy variados por lo que respecta a la extensión, al diseño y a las finalidades. El hecho es que los corpus han demostrado ser unas herramientas excelentes para muchos tipos de investigaciones; principalmente en el campo de la investigación lingüística porque, proporcionan bases mucho más reales para el estudio de las lenguas que los métodos intuitivos tradicionales. A partir de los corpus podemos disponer de bases muy provechosas para comparar diferentes variedades de una lengua o para explotar sus aspectos cuantitativos y probabilísticos. Los corpus han venido a dar un nuevo impulso a los estudios descriptivos de los diferentes aspectos de la lengua: prosodia, léxico, morfología, sintaxis, historia de la lengua, etc [24].

En este trabajo de utilizo un corpus de imágenes siendo de mucha ayuda a la hora de la traducción de texto a pictogramas, el corpus utilizado es el siguiente.

- Corpus de Pictogramas. Siendo este corpus la parte fundamental de esta investigación se contabilizaron más de 16,000 imágenes que representan números, letras, sentimientos, acciones etc. Se presenta una tabla 5.1 que describe cada tipo de pictograma *normal*, *concepto*, *ambiguo*. Del total de pictogramas, apro-

ximadamente 10,000 representan un concepto único, en la tabla ?? se detalla la cantidad de cada uno de ellos sin ser repetidos.

Tabla 5.1: Clasificación del total de pictogramas en el corpus.

Tipo	Cantidad
Normal	12496
Concepto	4056
Ambiguas	38
Total	16590

Tabla 5.2: Clasificación de pictogramas sin repetir, localizados en el corpus.

Tipo	Cantidad
Normal	7481
Concepto	2817
Ambiguas	11
Total	10309

## 5.2. Evaluación

La evaluación a dicha solución al sistema implementado se realizo calculando la medida de precision explicada anterior mente se ejecutaron distintas pruebas con diferentes texto traducidos a pictogramas como se aprecia en la tabla 5.3.

Como se puede observar se obtiene resultados aceptables al momento de la traducción teniendo un promedio superior al 80 % de efectividad.

Tabla 5.3: Porcentaje de efectividad en la traducción texto a pictogramas

Traducción			
Frase	Palabras por frase	Pictogramas encontrados	Porcentaje de efectividad
yo quiero una bicicleta	4	4	100 %
quiero tomar agua y comer carne	6	6	100 %
quiero un chocolate	3	3	100 %
voy a jugar con mi papa	6	4	67 %
voy a tomar el camión en la tarde	8	6	75 %
quiero ir con mi mama	5	4	80 %
quienes son tus papas	4	2	50 %
ya me voy a mi casa	6	5	84 %
quiero comer una hamburguesa con queso	6	5	84 %
quiero ver las noticias en la television	7	7	100 %
me acompañas a ir al baño	6	4	67 %

# Capítulo 6

## Conclusiones

Para el desarrollo de esta tesis se expuso la propuesta de solución para la traducción automática de texto a pictograma obteniendo el alcances de todos los objetivos planteados inicialmente el principal problema a tratar en esta investigación era solucionar la ambigüedad con el fin de obtener una traducción automática adecuada al contexto que se quiere expresar. Dicho problema fue solucionado por medio definiciones de cada palabra y por etiquetas POS (*PartOfSpeech* por sus siglas en inglés), se creo un etiquetado propio llamado *KarGus* con el cual se obtuvieron porcentajes de efectividad mayores al 70 %, logrando así oraciones con pictogramas completos.

Para esto se utilizo un corpus de pictogramas de origen español que contiene mas de 15 mil pictogramas de los cuales mas de 10 mil son únicos en esta tesis se utilizaron absolutamente todos los pictogramas dando así mejores traducciones y opciones de traducción para una misma oración.

Para dar solución al problema de ambigüedad a cada pictograma se le asigno su definición y su etiqueta POS, esto permite que al traducir un texto se identifique el contexto y a su vez se verifique si dicha oración tiene un orden gramatical y posteriormente identificar palabras ambiguas con el fin de presentar el pictograma mas adecuado a la oración.

La solución propuesta se implemento en un sistema web que permite el alcance a varios usuarios siendo esta una herramienta efectiva y de ayuda para la traducción de texto a pictogramas ayudando a personas a comunicarse con aquellas que presentan alguna discapacidad intelectual.

Como trabajo a futuro se pretende enriquecer el corpus de con mas imágenes adecuadas al idioma Español Mexicano para incrementar el nivel de efectividad al igual se considera realizar de manera mas personalizada la aplicación.

# Bibliografía

- [1] Doug Cutting, Julian Kupiec, Jan Pedersen, y Penelope Sibun. A practical part-of-speech tagger. En *Third Conference on Applied Natural Language Processing*. 1992.
- [2] Andrea Esuli y Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, págs. 617–624. ACM, 2005.
- [3] Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. En *IJCAI*, págs. 348–353. 2007.
- [4] Benjamin Fraser. El lenguaje visual innovador de pere joan: el pictograma analógico frente a la cultura digital en el cómic español contemporáneo. *Romance Studies*, págs. 1–16, 2018.
- [5] José Yuste Frías. Leer e interpretar la imagen para traducir. *Trabalhos em Linguística Aplicada*, 50(2):257–280, 2011.
- [6] Alexander Gelbukh. Procesamiento de lenguaje natural y sus aplicaciones. *Komputer Sapiens. Sociedad Mexicana de inteligencia aritifical*, 1, 2010.
- [7] España. Gobierno de Aragon. Arasaac (centro aragonees de tecnologias para la educacion). URL <<http://www.catedu.es/arasaac/aac.php>>.
- [8] Matías González, Hernán Sosa, y Adriana Elba Martin. Sistemas de comunicación no verbales. *Informes Científicos-Técnicos UNPA*, 6(2):30–56, 2014.

- 
- [9] Blanca Azucena Hidalgo Rodas y Mónica Alexandra López Rodríguez. *Los pictogramas en el desarrollo de la expresión oral y escrita de los niños y niñas de 4 a 5 años centro infantil parvulario politécnico del cantón Riobamba, provincia de Chimborazo en el periodo 2015-2016*. B.S. thesis, Riobamba, UNACH 2017, 2017.
- [10] S HRG. Committee on indian affairs united states senate. 2017.
- [11] Rolf G Hunt. Hash table implementation of an object repository. 2000. US Patent 6,154,747.
- [12] Geovanna Alexia Jaramillo Pazmiño. *los pictogramas y su relación con el desarrollo del lenguaje oral en los niños y niñas de 4 años de edad, del centro infantil verde manzana, de la ciudad de Quito en el período lectivo 2014-2015, lineamientos propositivos*. B.S. thesis, 2015.
- [13] Maryellen C MacDonald, Neal J Pearlmutter, y Mark S Seidenberg. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676, 1994.
- [14] James H Martin y Daniel Jurafsky. Speech and language processing. *International Edition*, 710, 2000.
- [15] Paula Andrea Benavides Ca non y Sandra Rodríguez Correa. Pln procesamiento del lenguaje natural en la recuperación de informción. 2007.
- [16] Eloy Martos Núñez y Gloria García Rivera. Estimulación lingüística a través de gráficos e imágenes. *Revista interuniversitaria de formación del profesorado*, (18):117–127, 1993.
- [17] María Helena Cortés Parazuelos. Una experiencia en clase: el chiste lingüístico. En *Tendencias actuales en la enseñanza del español como lengua extranjera I: actas del quinto Congreso Internacional de ASELE: Santander, 29, 30 de septiembre y 1 de octubre de 1994*, págs. 287–296. Asociación para la Enseñanza del Español como Lengua Extranjera, 1996.

- 
- [18] Christopher H Pratley, Erik J Rucker, David C Oliver, y Kentaro Urata. Method for converting a phonetic character string into the text of an asian language. 2002. US Patent 6,356,866.
- [19] José Luis Piñuel Raigada. Epistemología, metodología y técnicas del análisis de contenido. *Sociolinguistic Studies*, 3(1):1–42, 2002.
- [20] Cunalata Rochina y Adriana Paola. Los pictogramas favorecen la estructuración del lenguaje de los niños y niñas de 4 a 5 años del nivel de educación inicial de la unidad educativa ángel polibio Chávez, cantón guaranda, provincia bolívar, durante el periodo 2016–2017. 2017.
- [21] Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, y Frank Van Eynde. Extending a dutch text-to-pictograph converter to english and spanish. En *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, págs. 110–117. 2015.
- [22] Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, y Frank Van Eynde. Natural language generation from pictographs. En *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, págs. 71–75. 2015.
- [23] Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, y Frank Van Eynde. Simplified text-to-pictograph translation for people with intellectual disabilities. En *International Conference on Applications of Natural Language to Information Systems*, págs. 185–196. Springer, 2017.
- [24] Joan Torruella y Joaquim Llisterri. Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, págs. 45–77, 1999.
- [25] Teun Van Dijk. De la gramática del texto al análisis crítico del discurso. *BE-LIAR (Boletín de Estudios Lingüísticos Argentinos)*, 2(6):12–34, 1995.
- [26] Vincent Vandeghinste y Ineke Schuurman. Linking pictographs to synsets: Sclera2cornetto. En *LREC*, págs. 3404–3410. 2014.
- [27] Carmen Gierden Vega. La adquisición de un código cultural y lingüístico a través de la explotación didáctica de cómics y chistes gráficos en la clase de

- 
- alemán. *Encuentro: revista de investigación e innovación en la clase de idiomas*, (11):113–126, 1999.
- [28] Augusto Cortez Vasquez y Hugo Vega Huerta y Jaime Pariona Quispe y Ana Maria Huayna. Procesamiento de lenguaje natural. *Revista de investigacion de Sistemas e Informatica*, 6(2):45–54, 2014. ISSN 1816-3823.