# Procesamiento de lenguaje natural

Mg. Augusto Cortez Vásquez[1,2], Mg. Hugo Vega Huerta[1,2], Lic. Jaime Pariona Quispe[1]

[1]Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
[2]Facultad de Ingeniería
Universidad Ricardo Palma

cortez_augusto@yahoo.fr, hugovegahuerta@hotmail.com

**RESUMEN**

El artículo presenta el procesamiento de lenguaje natural mediante el modelado de los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas como son traducción, resúmenes de textos, recuperación de información, etc.

**Palabras clave:** lenguaje natural, procesamiento de lenguaje natural, análisis de lenguaje natural, lexicones

**ABSTRACT**

The present article present the processing of natural language by means of the shaped one of the cognitive processes that enter game the comprehension of the language to design systems that realize linguistic complex tasks since to be (translation, summaries of texts, recovery of information, etc.)

**Keywords:** Natural language, processing of natural language, analysis of natural language, lexicons

## I. INTRODUCCIÓN

La mayor parte del conocimiento científico es el resultado de muchos años de investigación, con frecuencia sobre temas aparentemente no relacionados. Y lo es mucho más en las ciencias de la computación, en donde el recurso más importante que posee la raza humana es información y conocimiento. En la época actual el uso de los recursos naturales, industriales y humanos depende del manejo eficiente de la información y conocimiento. Desde los tiempos antiguos hasta la actualidad, el conocimiento ha ido incrementándose a pasos agigantados en la forma de documentos, libros, artículos, y guardándose en diferentes formas: impresos, en forma electrónica (digital), con el advenimiento de las computadoras y el procesamiento del conocimiento el incremento ha sido mayor. Sin embargo, lo que es conocimiento para nosotros –los seres humanos– no lo es para las computadoras. La computadora almacena datos e información en archivos, puede copiar tal archivo, respaldarlo, transmitirlo, borrarlo, pero no puede buscar las respuestas a preguntas formuladas, hacer inferencias lógicas sobre su contenido, generalizar y resumirlo, es decir, hacer todo lo que las personas normalmente hacemos con el texto. Porque no lo puede entender.

Desde la perspectiva de la inteligencia artificial (IA), el estudio del lenguaje natural tiene dos objetivos:

**Objetivo 1:** Facilitar la comunicación con la computadora para que accedan a ella usuarios no especializados.

**Objetivo 2:** Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas (traducción, resúmenes de textos, recuperación de información, etc.)

Existen problemas en los que interesa fundamentalmente el primer objetivo. Esto se soluciona consiguiendo un intérprete para una clase de aplicaciones en un dominio restringido, que haga de traductor entre el computador y el usuario. El presente artículo se centra en el segundo objetivo, en el que se plantea el lenguaje como objeto de estudio, y la comprensión como un proceso complejo en que intervienen grandes cantidades de conocimiento de naturaleza diferente (morfología, sintaxis, semántica, pragmática) y mecanismos de tratamiento variados (de comparación, búsqueda, inferencia aproximada, deducción, etc.).

## II. GENERALIDADES

### Definición de lenguaje

Un lenguaje se puede definir de diferentes formas: desde el punto de vista funcional lingüístico se define como una función que expresa pensamientos y comunicaciones entre la gente. Esta función puede realizarse mediante signos escritos (escritura) o mediante señales y vocales (voz). Desde un punto de vista formal se define como un conjunto de frases, que generalmente es infinito y se forma con combinaciones de elementos tomados de un conjunto (usualmente infinito) llamado alfabeto, respetando un conjunto de reglas de formación (sintácticas o gramaticales) y de sentido (semánticas). Además de las características fundamentales del lenguaje debe considerarse que sea funcional, es decir, el lenguaje debe permitirnos expresar nuestras ideas. El lenguaje será bueno en la medida en que sea fácil de leer, fácil de entender y fácil de modificar. Lo mismo ocurre en los lenguajes formales[6].

Podemos distinguir entre dos clases de lenguajes: los lenguajes naturales (inglés, alemán, español, etc.) y lenguajes formales (matemático, lógico, programable etc.).

### Definición de lenguaje natural

Cuando queremos definir qué es lenguaje natural, nos hacemos la pregunta ¿Qué surgió primero las reglas gramaticales o el lenguaje? Un lenguaje natural es aquel que ha evolucionado con el tiempo para fines de comunicación humana, como el español o alemán [2]. Estos lenguajes continúan su evolución sin considerar la gramática, cualquier regla se desarrolla después de sucedido el hecho. En contraste, los lenguajes formales están definidos por reglas preestablecidas, y por tanto se rigen con todo rigor a ellas.

El lenguaje natural(LN) es el medio que utilizamos de manera cotidiana para establecer nuestra comunicación con las demás personas. El LN ha venido perfeccionándose a partir de la experiencia a tal punto que puede ser utilizado para analizar situaciones altamente complejas y razonar muy sutilmente. Los lenguajes naturales tienen un gran poder expresivo y su función y valor como una herramienta para razonamiento. Por otro lado, la sintaxis de un LN puede ser modelada fácilmente por un lenguaje formal, similar a los utilizados en las matemáticas y la lógica.

En un primer resumen, los lenguajes naturales se caracterizan por las siguientes propiedades:

1. Un lenguaje natural se define a partir de una gramática G, sin embargo, este se enriquece progresivamente modificando así también la gramática que la define. Esto dificulta la formalización de la definición de G.
2. Un LN tiene un gran poder expresivo debido a la riqueza del componente semántico (polisemántica). Esto dificulta aun más la formalización completa de su gramática.

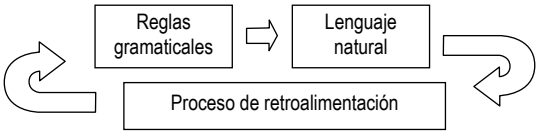



**Figura N.º 1.**

**Lengua y habla**

La lengua no es función del sujeto hablante, sino el producto que el individuo registra pasivamente. Nunca supone premeditación y la reflexión no interviene en ella más que para la actividad de clasificar.

El habla es el acto individual de voluntad y de inteligencia, ya que supone composición premeditada haciendo uso de la lengua. Cuando hablamos de la lengua y el habla, conviene distinguir:

A, Las combinaciones por lo que el sujeto hablante utiliza el código de la lengua con el objetivo de expresar sus ideas.

B. El mecanismo psicofísico que le permite exteriorizar esas combinaciones.

Al separar la lengua del habla se separa a la vez:

a. Lo que es social de lo que es individual

b. Lo que es esencial de lo que es accesorio

**Definición de lenguaje formal**

El lenguaje formal es aquel que el hombre ha desarrollado para expresar las situaciones que se dan en específico en cada área del conocimiento científico. Los lenguajes formales pueden ser utilizados para modelar una teoría de la mecánica, física, matemática, ingeniería eléctrica, o de otra naturaleza, con la ventaja de que en estos toda ambigüedad es eliminada. Revisten especial importancia los lenguajes de programación de computadoras, y estas se definen considerando un conjunto de componentes léxicos, reglas gramaticales y una delimitación semántica.

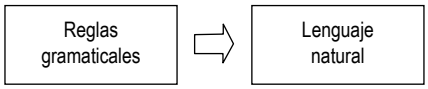



**Figura N.º 2.**

En resumen las características de los lenguajes formales son las siguientes:

1. Se desarrollan a partir de una gramática G preestablecida.
2. Componente semántico mínimo.
3. Posibilidad de incrementar el componente semántico de acuerdo con la teoría a formalizar.
4. La sintaxis produce oraciones no ambiguas.
5. Completa formalización y por esto, el potencial de la construcción computacional.

Antes de continuar con nuestro estudio del PLN, es importante el que estudiemos el concepto de lo que es un lenguaje de programación y las generaciones de estos para darnos una idea de cómo ha sido su evolución.

**Lenguaje de programación**

Un lenguaje de programación es un lenguaje formal definido como un conjunto de elementos (componentes léxicos) organizados a través de constructores (reglas gramaticales) que permiten escribir un programa y que éste sea entendido por el computador y pueda ser trasladado a computadores similares para su funcionamiento en otros sistemas. Un programa es una secuencia de instrucciones ordenadas correctamente que permiten realizar una tarea o trabajo específico. Un lenguaje de programación se basa en dos elementos muy importantes:

- **Sintaxis**: que se encarga del orden correcto de los componentes léxicos
- **Semántica:** se encarga de que cada "oración" del lenguaje de programación utilizado tenga un significado correcto.

## III. PROCESAMIENTO COMPUTACIONAL DEL LENGUAJE NATURAL(PLN)

Una de las tareas fundamentales de la Inteligencia Artificial (IA) es la manipulación de lenguajes naturales usando herramientas de computación, en esta, los len-

guajes de programación juegan un papel importante, ya que forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina. El PLN consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo ésta 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.

El uso del lenguaje natural (LN) en la comunicación hombre-máquina presenta a la vez una ventaja y un obstáculo con respecto a otros medios de comunicación.

**Ventaja**

Por un lado es una ventaja, en la medida en que el locutor no tiene que esforzarse para aprender el medio de comunicación a diferencia de otros medios de interacción como lo son los lenguajes de comando o las interfaces gráficas.

**Desventaja**

Su uso también tambien presenta limitaciones porque la computadora tiene una limitada comprensión del lenguaje. Por ejemplo, el usuario no puede hablar sobrentendidos, ni introducir nuevas palabras, ni construir sentidos derivados, tareas que se realizan espontáneamente cuando se utiliza el lenguaje natural. Realmente, lo que constituye en ventaja para la comunicación humana se convierte en problema a la hora de un tratamiento computacional, ya que implican conocimiento y procesos de razonamiento que aún no sabemos ni cómo caracterizarlos ni cómo formalizarlos.

**Aplicaciones del PLN**

Las aplicaciones del PLN son muy variadas, ya que su alcance es muy grande, algunas de las aplicaciones son:

- Traducción automática
- Recuperación de la información
- Extracción de Información y Resúmenes
- Resolución cooperativa de problemas
- Tutores inteligentes
- Reconocimiento de Voz

**Arquitectura de un sistema de PLN**

La arquitectura de un sistema de PLN se sustenta en una definición del LN por niveles: estos son : fonológico, morfológico, sintáctico, semántico, y pragmático.

**a. Nivel Fonológico:** trata de cómo las palabras se relacionan con los sonidos que representan.

**b. Nivel Morfológico:** trata de cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.

**c. Nivel Sintáctico:** trata de cómo las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.

**d. Nivel Semántico:** trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.

**e. Nivel Pragmático:** trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

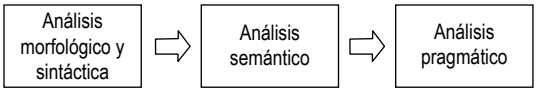



**Figura N.º 3.** Arquitectura de un Sistema de Procesamiento de Lenguaje Natural

La arquitectura del sistema de procesamiento del lenguaje natural muestra como la computadora interpreta y analizar las oraciones que le sean proporcionadas

La explicación de este sistema, es sencilla:

a. El usuario le expresa a la computadora que es lo que desea hacer.

b. La computadora analiza las oraciones proporcionadas, en el sentido morfológico y sintáctico, es decir, si las frases contienen palabras compuestas por morfemas y si la estructura de las oraciones es correcta. En esta etapa juegan un papel importante el analizador lexicográfico y el analizador sintáctico. El primero denominado scanner se encarga de identificar los componentes léxicos definidos a priori, el segundo denominado parser se encarga de verificar si se cumple un orden gramatical entre los elementos identificados por el scaner[2]

c. El siguiente paso es analizar las oraciones semánticamente, es decir saber cual es el significado de cada oración, y asignar el significado de estas a expresiones lógicas (cierto o falso).

d. Una vez realizado el paso anterior, ahora podemos hacer el análisis pragmático de la instrucción, es decir una vez analizadas las oraciones, ahora se analizan todas juntas, tomando en cuenta la situación de cada oración, analizando las oraciones anteriores, una vez realizado este paso, la computadora ya sabe que es lo que va a hacer, es decir, ya tiene la expresión final.

e. Una vez obtenida la expresión final, el siguiente paso es la ejecución de esta, para obtener así el resultado y poder proporcionárselo al usuario.

**Sintaxis y Gramática**

La sintaxis se define como la disposición de palabras en una oración para mostrar su relación. Describe la secuencia de símbolos que constituyen programas validos[3,4]. En un lenguaje de programación convencional la frase **a = b + c** representa una secuencia valida de símbolos, pero **c = b a +** no lo es. Esto se justifica, dado que en una sentencia de asignación el lado izquierdo del operador de asignación debe ser un identificador y el lado derecho debe haber una expresión aritmética valida. La sintaxis suministra información significativa que se necesita para entender un programa y proporciona información imprescindible para la traducción del programa fuente a un programa objeto[11]. La sintaxis muestra al hablante la forma como debe escribir buenos oraciones. La sintaxis es más útil al usuario del lenguaje que al sistema de PLN.

Una gramática G es un modelo linguístico-matematico que describe el orden sintáctico que den cumplir las frases bien formadas de un lenguaje[1,2]. Una gramática se define formalmente de la siguiente forma:

G =( $V_t$ , $V_N$, P, S) donde:

$V_T$ : conjunto finito de símbolos terminales del lenguaje

$V_N$ : conjunto finito de símbolos no terminales

P : conjunto finito de reglas de producción

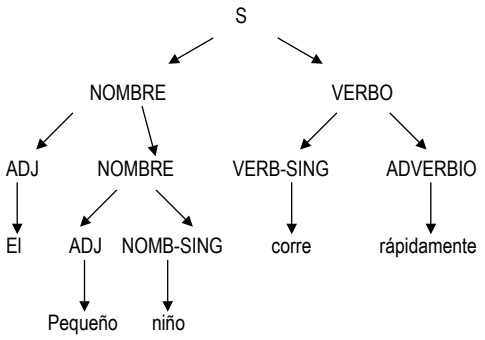
S : Símbolo distinguido o axioma inicial a partir del cual se reconocerán las secuencias de L aplicando sucesivamente las reglas de producción.

Consideremos la siguiente gramática G ($V_N$, $V_T$, S, P) donde

$V_N$ = {S, NOMBRE, VERBO, ADJ, NOMB-SING, VERBO-SING,ADVERBIO}

$V_T$ = {El, La, Los, Las, Pequeño, traviesa, niño, niña, estudia, corre, juega, salta}

P = {S → NOMBRE VERBO
NOMBRE → ADJ NOMBRE
NOMBRE → ADJ NOMB-SING
VERBO → VERB-SING ADVERBIO
ADJ → El /La /Los /Las /Ellos /Ellas
ADJ → Pequeño /traviesa/ inquieto
NOMB-SING → niño / niña
VERB-SING → estudia / corre / juega /salta
ADVERVIO → rápidamente / despacio / mucho
}




luego w = 'El Pequeño niño corre rápidamente' Є L(G)

Durante el análisis sintáctico, se realizan derivaciones (de reglas gramaticales) a partir de un símbolo distinguido, para verificar si una frase pertenece al lenguaje definido por la gramática. A este proceso para determinar si es gramaticalmente correcta o no se le conoce como análisis sintáctico (*parsing*). Los árboles de análisis sintáctico muestran la sintaxis concreta de un lenguaje [3,6]. Sin embargo, para aplicar esta gramática de forma mecánica y automatizada a una oración, es necesario contar con un lexicón que ofrezca información al analizador sintáctico (*parser*) sobre las categorías gramaticales que están asociadas a las palabras que aparecen en la oración que se desea analizar. El análisis semántico es a la vez la fase medular de las instrucciones. Aquí se procesan las estructuras sintácticas reconocidas por el analizador sintáctico. Un analizador semántico puede estar constituido por un conjunto de

analizadores semánticos mas pequeños. Cada uno de los cuales, maneja un tipo particular de construcción. Estos interactúan entre ellos mismos a través de información que se guarda en estructuras de datos.

Consideremos otra gramática G ($V_N$, $V_T$, S, P) donde

$V_N$ = {A,S,P}

$V_T$ {s, v, p, y}

Una oración tiene la forma S V P s: sujeto, v: verbo p: predicado

Definimos las producciones

$A \rightarrow S\ V\ P$

$S \rightarrow s$ / S y s

$V \rightarrow v$ / V y v

$P \rightarrow p$ / P y P

Donde

s : corresponde a sujeto: Juan, el, ellos, aquellos, etc.

v : corresponde a un verbo: jugar, estudiar, saltar, etc.

p : corresponde a un predicado: poco, mucho, despacio, etc.

A → SVP se remplaza por $A \rightarrow SA_1$, $A_1 \rightarrow V\ P$

S → s/S y s se remplaza por $S \rightarrow s$, $S \rightarrow SA_2$, $A_2 \rightarrow YS$

V → v/V y v se remplaza por $V \rightarrow v$, $V \rightarrow VA_3$, $A_3 \rightarrow YV$, $Y \rightarrow y$

P → p/P y p se remplaza por $P \rightarrow p$, $P \rightarrow P\ A_4$, $A_4 \rightarrow YP$

Ejemplo: María Esther y Karla saltan y cantan contentas y sonrientes.

**Problema del procesamiento de lenguaje natural**

La principal dificultad en los procesos de recuperación de información mediante lenguajes formales no es de índole técnica sino psicológica: entender cuál es la necesidad real del usuario, cual es la correcta formulación de su pregunta o necesidad. La dirección más prometedora de resolver este problema es el uso de lenguaje natural. Sin embargo, uno de los grandes problemas del PLN se produce cuando una expresión en LN posee más de una interpretación, es decir, cuando en el lenguaje de destino se le pueden asignar dos o más expresiones distintas. Este problema de la ambigüedad se presenta en todos los niveles del lenguaje, sin excepción. Ejemplo:

"Hay alguien en la puerta, que te quiere hablar"

" Hay alguien, en la puerta que te quiere hablar"

No está claro, si el predicado "te quiere hablar" se adjudica a "alguien" o a "la puerta", sabemos que la puertas no hablan, por tanto deducimos que es a alguien. Pero esto no lo puede deducir la máquina, a no ser que esté enterada de lo que hacen o no hacen las puertas. En apariencia este problema es demasiado sencillo, pero en realidad, es uno de los más complicados y que más complicaciones ha dado para que el PLN pueda desarrollarse por completo, ya que al presentarse en todos los niveles del lenguaje, se tienen que desarrollar programas (lenguaje formal) para solucionarlos en cada caso.

**El PLN en los Sistemas Multimedia y Expertos: Tutores Inteligentes(TI)**

La informática ha evolucionado desde sus inicios, considerando siempre aspectos del comportamiento del usuario en relación con el tratamiento de la información. Es por eso que ha incorporado textos, imágenes y sonido a las estaciones de trabajos actuales, al tiempo que éstos aumentan su capacidad.

Los sistemas multimedia incluyen:

1. Entornos visuales
2. Autopistas de información
3. Ratón
4. Programación interactiva
5. Realidad Virtual
6. Hipertexto
7. Sonido

La multimedia combina el hipertexto con el sonido. Estas uniones de imágenes, texto y sonidos necesitan una filosofía del conocimiento que fundamente su función interna dentro de la comunicación de conocimientos. Existe una comunicación sistema-usuario que se da a través de un lenguaje natural que se ve afectado grandemente por el conocimiento que un interlocutor tiene del otro y por el contexto o entorno donde el diálogo tiene lugar.

## IV. EL LEXICÓN EN EL ÁMBITO DE LA PSICOLINGÜÍSTICA: EL LEXICÓN MENTAL

La complejidad de la memoria léxica ha fascinado a muchos psicolingüistas, sobre todo la forma cómo éste se organiza en la memoria de un hablante para su acceso y uso inmediato, a tal punto que han propuesto diferentes métodos para explorar y analizar los procesos cognitivos que se producen en su uso. El hecho de que un hablante pueda acceder en milésimas de segundo a una cantidad ingente de vocabulario almacenado en su memoria, tanto en procesos de producción como de comprensión, es una prueba fehaciente de que el lexicón mental está organizado y estructurado de modo que posibilita el acceso inmediato. En la dimensión de la psicolingüística, se define el lenguaje interiorizado, como una actividad mental interna. La lingüística atiende a reglas y estructuras de la gramática de una lengua. La psicolingüística estudia procesos y representaciones implicadas en la comprensión, adquisición y producción del lenguaje[11].

De entre los modelos explicativos del acceso y procesamiento de la información léxica debemos destacar los siguientes:

**a. Modelos de activación.** Cada elemento léxico tiene asociado un logogen que permanece activado durante todo el proceso de recuperación de una determinada unidad léxica. Activa las palabras que se corresponden con la información sensorial

**b. Modelos autónomos.** El acceso léxico se realiza solo por medio de información sensorial, sin que haya interacción con otros componentes del sistema cognitivo.

**c. Modelos modulares.** Sostiene la existencia de módulos separados que contienen información fonológica, ortográfica, sintáctica y semántica de las palabras. Experimentos realizados con pacientes afásicos o con disfunciones en el habla favorecen la hipótesis de la modularidad en la estructura del lexicón mental, ya que en casos de daños cerebrales el acceso a la información fonológica, ortográfica, sintáctica y semántica de las palabras puede verse afectada de manera independiente.

**d. Los modelos de redes semánticas.** Propuestos por Collins y Quillian, intentan describir y explicar cómo la información se almacena de modo "económico" en el cerebro en forma de redes, en las que se incorporan dos tipos básicos de relaciones: relaciones "IS-A" y relaciones "HAS-A", (es decir, relaciones de hiperonimia y relaciones de meronimia), aunque otros tipos de relaciones semánticas, tales como sinonimia o la antonimia se consideran también necesarias para describir la estructura del lexicón mental.

Investigaciones realizadas acerca del aprendizaje y crecimiento de vocabulario en niños de edades entre seis y ocho años, han demostrado que a esa edad, la "perceptibilidad léxica" está muy desarrollada y que los niños son especialmente perceptivos a las palabras nuevas, pudiendo deducir su significado del contexto en el que las oyen, y llegando a aprender una media de 21 palabras nuevas cada día. En este proceso de aprendizaje, el niño debe primero asignar la palabra nueva a una categoría semántica, y debe aprender a distinguirla de las demás palabras asignadas a la misma categoría, de modo que parece imposible que los niños aprendan un número tan elevado de palabras, en un periodo tan corto de tiempo, a no ser que las organicen en su mente estructurándolas de algún modo a través de tipos, y la mayoría de los experimentos señalan hacia la organización en campos léxicos.

## V. EL LEXICÓN EN EL PROCESAMIENTO DE LENGUAJE NATURAL: LA LEXICOGRAFÍA COMPUTACIONAL

Actualmente, en el ámbito computacional, los lexicones se consideran la base fundamental en la construcción de sistemas computacionales que posibiliten la interacción entre la máquina y el hombre. No se pueden construir sistemas de procesamiento de lenguaje natural que sean lo suficientemente robustos como para ocuparse de problemas del "mundo real", sin antes diseñar lexicones de gran magnitud que contengan información léxica detallada[16,18] .

Se distinguen dos grandes ámbitos de investigación en lo referente a los lexicones computacionales: el de la ***adquisición*** y el de ***representación*** de conocimiento léxico.

### Adquisición de conocimiento léxico

El gran problema al que se enfrentan en el diseño de sistemas de lenguaje natural a gran escala, es el gran número de unidades léxicas de las lenguas naturales, así como a la constante incursión de palabras nuevas o nuevas acepciones de palabras existentes.

La adquisición de la información léxica necesaria para lexicones computacionales plantea serios problemas, tanto en lo que se refiere a la efectividad de los diferentes métodos que se han empleado como a la inversión de tiempo, dinero y recursos humanos y computacionales que estos métodos requieren[12].

Se puede considerar que existen tres métodos o fuentes principales para la adquisición de conocimiento léxico:

1. Adquisición manual de información léxica
2. Diccionarios en formato magnético (MRDs)
3. Los córpora textuales informatizados

Los tres métodos plantean ventajas y desventajas, tanto en lo que se refiere a los recursos que requieren como a la efectividad que han demostrado hasta ahora.

Aunque en principio las fuentes electrónicas pueden aportar una gran cantidad de información lingüística muy valiosa, que puede servir como punto de partida para la creación de una base de datos léxica, en la práctica es difícil aprovechar toda la información que esas fuentes electrónicas contienen. Una de las dificultades, y quizás la principal, es que los diccionarios están diseñados por humanos (y no máquinas ) para ser usados por humanos. Los usuarios (humanos) son hablantes nativos de una lengua, que conocen el contexto de lo que se está hablando, y saben implícitamente, cómo está estructurado el lexicón de su lengua. Los MRDs, en muchas ocasiones, son elaborados por lexicografos, quienes explotan el conocimiento lingüístico de sus usuarios potenciales, de modo que las entradas de un diccionario contienen solo la información necesaria para que un hablante de una lengua sea capaz de conectarla con su conocimiento lingüístico general[15].

Karen Sparck-Jones demostró en un estudio realizado que los diccionarios deben contener un componente de circularidad, ya que cada palabra usada en las definiciones ha de ser, a su vez, definida en el diccionario. Algunas de estas circularidades mantienen una distancia semántica reducida, como por ejemplo las definiciones mutuas de “bueno” y “excelente”, y son por tanto fáciles de observar y asimilar por un lector humano, pero son muy difíciles de localizar a nivel formal lo cual dificulta la labor de extracción de información de las definiciones.

El **lexicón** se considera como un “diccionario mental” en el que se registran las palabras que conoce un hablante. Este “diccionario” especifica los rasgos característicos de los componentes léxicos (palabras y morfemas), como irregularidades morfológicas, requerimientos sobre alomorfos, información pragmática, etc. Un símbolo **alomorfo** se refiere a cada uno de las diferentes formas fonológicas que puede tener un morfema abstracto. Estrictamente la realización fonológica concreta de un morfema se llama morfo, si existe más de un morfo para el mismo morfema entonces usamos el término alomorfo.

Algunos modelos gramaticales formales basan la generación de oraciones en el procesamiento de los rasgos de las unidades del lexicón. En estos modelos, el lexicón no es parte de la gramática, sino que proyecta sus rasgos a través de mecanismos inherentes a las gramáticas. La finalidad fundamental del procesamiento de lenguaje natural es la automatización de los procesos lingüísticos, tales como la comprensión, producción o adquisición de una lengua, tareas que los usuarios de una lengua realizan fluida y naturalmente. Esto hace converger intereses de varias disciplinas como son lingüistas computacionales, psicolingüistas, informáticos e ingenieros de sistemas. Todos ellos, desde diferentes perspectivas teóricas y prácticas, intentan desarrollar una teoría que sea totalmente explícita (y por tanto automatizable) de los procesos lingüísticos.

La mayoría de los sistemas de procesamiento de lenguaje natural adoptan un enfoque denominados “basado en el conocimiento” (*knowledge-based*), ya que para llevar a cabo la tarea para la que están diseñados, necesitan incorporar conocimiento lingüístico explícito, junto con otros tipos de conocimiento de carácter más general. Por ejemplo, un sistema que convierta un texto en su correspondiente cadena hablada, necesita “conocimiento” sobre la pronunciación de las letras, así como de las palabras individuales que no siguen las reglas generales. También precisa conocimiento sobre los patrones rítmicos de acentuación y de cómo la organización sintáctica afecta la entonación y prosodia. Atendiendo estas consideraciones, con el objetivo de consensuar en la investigación sobre el PLN, se ha dividido su estudio en subsistemas , en relación con los niveles presentados en la arquitectura de un sistema de PLN, identificando cinco tipos de conocimiento:

| Conocimiento fonológico | Conocimiento morfológico | Conocimiento sintáctico: | Conocimiento semántico: | Conocimiento pragmático |
|---|---|---|---|---|
| información sobre el sistema de sonidos y la estructura de las palabras y las expresiones, los patrones de acentuación, la entonación, etc. | información sobre la estructura de las palabras; por ejemplo, que los fonemas /s/ y /z/ se añaden en inglés a los nombres para formar el plural. | información sobre las reglas sintácticas y/o gramaticales. | información sobre el significado que se da a las diversas construcciones sintácticas y de cómo esos significados se combinan para formar el significado de las oraciones. | información central en muchas tareas específicas como por ejemplo, la recuperación de los referentes de los pronombres, las intenciones comunicativas que subyacen en una frase en particular, el análisis de las presuposiciones del hablante. |

La noción de sistema o estructura surge como reacción al atomismo lingüístico, en la que se entiende el lenguaje de manera aislada, no en términos de relaciones de unos componentes con los otros. Por ejemplo, un sistema fonológico no es la suma mecánica de los fonemas aislados, sino un todo orgánico cuyos fonemas son los miembros y cuya estructura está sujeta a ciertas leyes. Lo importante no son los elementos constitutivos, ni su totalidad resultante, sino las relaciones que expresan en términos de leyes.

Cada uno de estos cinco tipos de conocimiento puede ser caracterizado por medio de un conjunto de reglas. Por ejemplo, es una regla de tipo sintáctico en español que las oraciones tengan la siguiente estructura: sujeto + verbo+ predicado, ejemplo “Juan estudia mucho”. El lexicon debe explicitar este tipo de particularidades.

El lexicón debe adaptarse a la gramática diseñada, pero ambos tendrían que ser extendidos cada vez que se introdujeran reglas nuevas en la gramática o se añadieran palabras al lexicón. Tradicionalmente en español se han reconocido verbos predicativos (transitivos e intransitivos) y tres verbos copulativos: *ser*,*estar*, y *parecer* y estos nunca pueden llevar complemento directo; en cambio, llevan un complemento llamado atributo, que suele ser un sustantivo o adjetivo representando un estado o cualidad del sujeto. Si añadimos, por ejemplo, un verbo no copulativo, como solitaria, necesitaríamos hacer una distinción entre diferentes tipos de verbos, tanto en la gramática como en el lexicón, para evitar que se generen oraciones incorrectas. Esto demuestra la necesidad de que en cualquier sistema de procesamiento de lenguaje natural exista una gran interconexión entre las reglas generales que se incorporan a la gramática y la información incluida en las entradas del lexicón, ya que el lexicón deberá aportar toda la información que no sea predecible de las reglas, y deberá “rellenar” estas reglas de modo que funcionen correctamente.

El lexicón también tiene que incluir otros tipos de información no derivable de reglas, como por ejemplo, información idiosincrática, de pronunciación, que en caso del inglés por ejemplo se considera normalmente como un aspecto lingüístico que no se puede derivar del significado de las palabras o de su forma morfológica.

**Agradecimientos**

El presente trabajo se desarrolla en el marco del proyecto de investigación, financiado parcialmente por el Vicerrectorado de Investigación de la Universidad Nacional Mayor de San Marcos.

**Trabajos futuros**

A partir del conocimiento generado en disciplinas como la informatica y la lingüística computacional, se están desarrollando sistemas para la confección de resúmenes y la indización automática. Este tipo de investigaciones se lleva practicando desde hace tiempo, y se comienza a recoger los frutos de años de inspección, por lo que se debe permanecer atentos a su evolución. El procesamiento del lenguaje natural es una labor compleja, no exento de dificultad para los lingüísticas que deben adquirir la instrumentación de los informáticos, y para los informáticos, ya que deben hacer suyos conocimientos lingüísticos.

## VI. CONCLUSIÓN

1. El lenguaje natural (LN) nos permite el designar las cosas actuales y razonar acerca de ellas, fue desarrollado y organizado a partir de la experiencia humana y puede ser utilizado para analizar situaciones altamente complejas y razonar muy sutilmente.
2. Los lenguajes de programación (LP) son un tipo muy limitado de lenguaje natural, orientados básicamente a la manipulación de datos e información discreta, pero no son suficientes para la comunicación integral que incluya la totalidad de los aspectos semánticos y pragmáticos.
3. El procesamiento de lenguaje natural (PLN) consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo esta 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.

   Los lexicones son una parte importante del procesamiento de lenguaje natural y debe contener información fonológica, morfológica, sintáctica, semántica y pragmática, pero además esta información debe ser estructurada de forma que permita su reutilización para diversas tareas.
4. El lexicón también tiene que incluir otros tipos de información que considere aspectos de orden idiosincrática, de pronunciación, y toda información que no se puede derivar del significado de las palabras o de su forma morfológica.

## VII. BIBLIOGRAFÍA

[1] [AHO 1990] Aho A.,Sethi,Ullman *Compiladores, principios, técnicas y herramientas*; Addison-Wesley 1990, Wilmington-Delaware EUA.

[2] [BROOKSHEAR 1993] BROOKSHEAR J. Glean. *Teoría de la computación* Addisson Wesley iberoamericana Wilmington Delaware 1993.

[3] [CORTEZ 2002] Cortez Vásquez, Augusto. *Lenguajes y compiladores*, UNMSM EAPIS 2002.

[4] [HOPCROFT 1993] Hopcroft Jhon, Ullman Jeffrey. *Introducción a la teoría de autómatas*. Edit. CECSA 1993.

[5] [PRATT 1988] Terrence W. Pratt. *Lenguajes de programación, Diseño e implementación*; Prentice Hall Hispanoamericana 1988.

[6] [SETHI 1992] SETHI, Ravi *Lenguajes de programación, Conceptos y Constructores*; Addison-Wesley, 1992.

[7] [TEUFEL 1990] Teufel-Smithd-Teufel. *Compiladores, Conceptos fundamentales*; Addison-Wesley, 1990.

[8] La construcción del WordNet 3.0 en español, ANA FERNÁNDEZ MONTRAVETA. Universitad Autónoma de Barcelona GLORIA VÁZQUEZ.

[9] Letch, Charley. *Información Tsunami: Un futurista mira en retrospectiva*, Primera Edición, Editorial. Limusa, Colección Megabyte, México D.F., 1992

[10] http://delta.cs.cinvestav.mx/red/logica/node3.html

[11] http://cic2.iimas.unam.mx/~villasen/protocolo-proy-CONACYT.html

[12] http://www3.uniovi.es/~Psi/REMA/v1n1/a4/p1.html

[13] http://www.dcc.uchile.cl/~cc20a/contenidos/clase05

[14] http://www.lawebdelprogramador.com/

[15] http://es.thefreedictionary.com/lexicones [2010]

[16] http://elies.rediris.es/elies19/cap3443.html

[17] http://elies.rediris.es/elies9/2.htm