# 2 Estado del arte

Este capítulo tiene como finalidad enumerar y explicar las bases técnicas que han sido necesarias para el desarrollo del proyecto.

En el primer subcapítulo se va a tratar la Ingeniería Lingüística, su definición, las necesidades de esta ingeniería y el lenguaje y los diversos problemas de la lingüística en el procesamiento del lenguaje natural.

El procesamiento del lenguaje natural es una de las partes fundamentales de este proyecto, por lo que se explicará su arquitectura típica, la segmentación, detección de entidades y etiquetados y análisis sintáctico y morfológico.

Por último, puesto que el desarrollo del sistema se ha realizado en el lenguaje estructurado XML, también se dedicará un subcapítulo de este estado del arte a explicar las bases de este lenguaje.

En el desarrollo de estos subcapítulos se empleará el uso de diversos ejemplos para facilitar la comprensión de los temas descritos.

## 2.1 Ingeniería lingüística

El término de Ingeniería lingüística abarca un amplio espectro de actividades que suelen englobarse dentro de lo que se ha denominado “las industrias de la lengua”. Comprende una serie de técnicas relacionadas con el tratamiento informático del lenguaje, en este caso del lenguaje natural. El lenguaje natural es el lenguaje que usamos los humanos para comunicarnos y expresarnos.

Este tratamiento informático del lenguaje solía dividirse entre las técnicas que se aplicaban al lenguaje hablado y las propias del procesamiento del texto escrito, pero existe cada vez una mayor convergencia entre ambos métodos. Lo que se persigue es mejorar la interoperabilidad de los sistemas informáticos. Se consigue perfeccionar la interacción de los usuarios con los sistemas informáticos, iteración hombre-máquina más adecuado; y además, que el sistema sea capaz de asimilar, seleccionar, utilizar y presentar la información en función de las necesidades de la aplicación que se desee darle.

### 2.1.1 Definición

La Ingeniería Lingüística según la Wikipedia [1] “es una disciplina también denominada Informática aplicada a la Lingüística e incluso Tecnología del Lenguaje, que, a su vez, tiene carácter multidisciplinar. La ingeniería lingüística se aprovecha del conocimiento desarrollado en el marco informático del procesamiento del lenguaje natural y del marco lingüístico nutrido por las disciplinas de la traducción, de la terminología y de la lingüística computacional, tanto en sus vertientes teóricas como aplicadas. ”

En definitiva, se podría definir la Ingeniería Lingüística como la tecnología encargada del lenguaje natural, en cuanto a procesamiento y lo que ello conlleva, así como sus posibles aplicaciones.

### 2.1.2 El lenguaje y las necesidades de la ingeniería lingüística

El lenguaje puede ser entendido como un recurso que hace posible la comunicación. En el caso de los seres humanos, es una herramienta que se encuentra extremadamente desarrollada, que brinda la posibilidad al hombre de selección, citar, coordinar y combinar conceptos de diversa complejidad.

Existen numerosas maneras del lenguaje. El concepto de lengua natural o lenguaje natural, describe a una modalidad lingüística o tipo de lenguaje que el hombre desarrolla con el propósito de comunicarse con su entorno. Esta herramienta, según se advierte al analizar sus particularidades, posee sintaxis y tiene su base en los preceptos de optimización y economía. Además el leguaje ayuda a crear una cultura y se transmite de generación en generación, es decir, está ligado al concepto de sociedad.

Se vive en un mundo en el que el acceso a la información es cada vez más una necesidad, esta necesidad se ha visto impulsada por el desarrollo de las nuevas tecnologías. Éstas están al alcance de cualquier usuario que esté dispuesto a usarlas y ofrecen una gran cantidad de datos y de modos de interacción, gracias a Internet podemos hablar con cualquier persona sin importar la distancia, comprar, tener información de todo tipo. Todo esto lo recoge el concepto de sociedad de la información, aquella en la cual las tecnologías facilitan la creación, distribución y manipulación de la información juegan un papel importante en las actividades sociales, culturales y económicas.

Esta sociedad de la información tiene sus inconvenientes, entre otros el exceso de información que existe actualmente en la red, en estos tiempos acceder a todo tipo de información es relativamente fácil, pero se ha de disponer de los medios para obtener exclusivamente la información de lo que se desea saber de esa información.

Debido a este nuevo entorno que se ha creado, las nuevas tecnologías deben dar un acceso a la información que sea rápido, sencillo y eficaz.

Así el desarrollo de aplicaciones se tendrá que enfocar en crear programas que procesen el lenguaje, para ello se tendrán que generar algoritmos y funcionalidades. Todo ello teniendo en cuenta el gran volumen de información que se va a tratar y contemplar los posibles errores que puedan introducir los usuarios.

A consecuencia de todo lo expuesto surge la problemática de la Recuperación de Datos, implica que esta ingeniería desarrolle aplicaciones capaces de analizar los recursos de información disponibles localizando e interpretando los datos para la resolución del problema que plantea. Es decir, recuperar la información textual que satisfaga la necesidad de información del usuario. Facilitando la descripción del contenido de documentos así como para representar las consultas formuladas por el usuario.

**Procesamiento del lenguaje natural. Niveles del lenguaje**

El lenguaje natural, como se ha descrito anteriormente, es la modalidad lingüística que el hombre ha desarrollado para comunicarse con su entorno; por tanto, dentro de la ingeniería lingüística existe una especialidad que se encarga del estudio de este lenguaje. Las labores de esta especialidad son la de construir sistemas y mecanismos que permitan o faciliten la comunicación entre personas y máquinas para así facilitar la búsqueda de información. Uno de los grandes retos de la informática es el desarrollo de ordenadores que sean capaces de entender el lenguaje natural.

El lenguaje natural, o mejor dicho, la sintaxis del lenguaje natural se corresponde a un lenguaje formal, similar a los lenguajes lógicos y matemáticos, con lo que se permite que pueda ser modelado.

Para estudiar la forma en la que se estructura el lenguaje natural se utiliza lo que se denominan, tradicionalmente, niveles del lenguaje. En total existen cinco niveles del lenguaje que se detallan a continuación:

- **Fonología**
- **Morfología**
- **Sintaxis**
- **Semántica**
- **Pragmática**

**Fonología**

La fonología [2] es una disciplina basada en el estudio de los fonemas o segmentos mínimos de la corriente fónica que tienen entidad como elementos del sistema lingüístico, teniendo en cuenta su valor distintivo y funcional. El número de fonemas de una lengua oscila entre veinte y cuarenta.

No hay que confundir la fonología con la fonética, ya que esta última se encarga de estudiar la naturaleza acústica de los sonidos, es decir, la articulación de los sonidos. Mientras que la fonología describe el modo en que los sonidos funcionan, en una lengua en particular o en las lenguas en general, en un nivel abstracto o mental.

Dentro de este campo hay que definir lo que son los pares mínimos, se trata de palabras distintas con significado diferente y que sólo varían en un sonido. Un par de ejemplos para entender los pares mínimos pueden ser: *casa* y *masa*, o *boca* y *roca*.

Como hemos visto los sonidos que componen una palabra son las unidades mínimas que la hacen diferente de otra, es decir, que hay unidades mínimas que diferencian los significados, los fonemas [3].

Los fonemas se definen siguiendo unas normas físicas y articulatorias, en función de su carácter sonoro o sordo. Como ejemplos de tipos de sonidos que se pueden encontrar en una lengua están: bilabial, nasal, consonántico, etc.

Entre los principales rasgos fonéticos que se tienen en cuenta para distinguir fonemas aparecen su **consonanticidad**, su **silabicidad**, su **sonoranticidad**, su **sonoridad** y **aspiración**, su **modo de articulación** y su **punto** o **lugar de articulación**.

Para las aplicaciones de audio o vídeo, como son: los programas de escritura predictiva, programas que generan audio a partir de texto, etc. la fonología es una parte del lenguaje que hay que tener en cuenta. Sobre todo, teniendo en cuenta la cantidad de ambigüedades que pueden aparecer debido al tipo de entonación, letras mudas, fonemas parecidos entre palabras, etc.

Los problemas en fonética computacional están conectados al desarrollo de sistemas de análisis y síntesis del habla. Aun cuando hay sistemas de reconocimiento de voz, el porcentaje de palabras identificadas correctamente es todavía bastante bajo. Entre sistemas de generación de voz hay mucho más progreso, basados en síntesis compilativas, aunque su área de aplicación es bastante restringida.

**Morfología**

La morfología [4] es una rama de la lingüística que se encarga de estudiar la estructura interna de las palabras.

Trata de delimitar, definir y clasificar las clases de palabras a las que da lugar (morfología flexiva) así como la formación de nuevas palabras (morfología léxica).

La historia de la morfología data del siglo XIX y únicamente en sus inicios trataba la forma de las palabras, actualmente su acepción es estudiar fenómenos más complejos que simplemente la forma.

Un ejemplo claro de morfemas es el siguiente: pato, la palabra se forma con dos monemas.

El lexema (raíz de la palabra), es *pat* y los morfemas (prefijos, sufijos o desinencias) son {*-o -a -os -as*} dando lugar a las palabras: *pato, pata, patos, patas*.

Por tanto, del anterior ejemplo se puede sacar que dado un conjunto de palabras que comparten un mismo lexema, éste se puede denominar raíz.

Dentro del campo de las aplicaciones informáticas la morfología se usa para descomponer y etiquetar las palabras para así hacer el análisis de una palabra. Con esto se consiguen programas tales como correctores ortográficos o gramaticales. Los problemas de la morfología computacional están relacionados con el desarrollo de los sistemas de análisis y síntesis automático morfológico. El desarrollo de tales módulos es bastante difícil porque hay que hacer grandes diccionarios de raíces, en general existe la metodología para tal desarrollo y existen sistemas funcionando para muchos idiomas.

**Sintaxis**

La sintaxis [5] es la parte de la gramática que estudia la forma en que se combinan y relacionan las palabras para formar secuencias mayores, cláusulas y oraciones.

Hay que destacar que el análisis sintáctico de una oración supone la búsqueda del verbo conjugado y así llegar a distinguir entre el sintagma sujeto y el sintagma predicado. Una forma sencilla de identificar estos sintagmas es una vez ubicado el verbo, preguntar quién o qué realiza la acción. La respuesta a esa pregunta, es el sujeto mientras que el resto es el predicado.

La sintaxis es la parte de la gramática que estudia las reglas que gobiernan o rigen la combinatoria de constituyentes sintácticos y la formación de unidades superiores a estos, como los sintagmas y oraciones gramaticales. La sintaxis, por tanto, estudia las formas en que se combinan las palabras, así como las relaciones sintagmáticas y paradigmáticas existentes entre ellas.

La sintaxis computacional debe tener métodos automáticos para análisis y síntesis, es decir, para construir la estructura de la frase por la frase, o generar la frase con base en su estructura. El desarrollo de los generadores es una tarea mucho más fácil, y está más o menos claro que algoritmos son necesarios en estos sistemas. Al contrario, en el desarrollo de los analizadores sintácticos, todavía es un problema, especialmente para los idiomas que no tienen un orden de palabras fijos, como en el español, por ello las teorías basadas en inglés no son fáciles da adoptar al español.

**Semántica**

Semántica [6] (proviene del griego *semantikos***,** "lo que tiene significado"), es el estudio del significado de los signos lingüísticos (palabras, expresiones y oraciones).

Para ello se tienen que estudiar qué significa para los hablantes, cómo los designan, y también cómo los interpretan los oyentes.

Todo signo tiene dos vertientes: el significante o parte material del mismo y el significado o imagen que mentalmente genera el significante. Además, hay que distinguir otro elemento que es el referente o elemento real, al que se refieren tanto significado como significante. No es lo mismo la palabra que designa un referente que el referente mismo.

El significado o imagen mental está compuesto por una serie de rasgos que todos los hablantes de una lengua asocian de una manera general a un significante. No obstante, este significado tiene dos componentes:

- **Denotación.** Son los rasgos objetivos. Es el significado concreto de una palabra fuera de contexto. Constituyen el núcleo semántico fundamental. Comunes para todos los hablantes. Es el significado que se puede encontrar en el diccionario.
- **Connotación**. Son los rasgos subjetivos. Son las significaciones que lleva añadidas una palabra de manera subjetiva. De modo que dependiendo de los hablantes, época o lugar no cobre un significado distinto. Como por ejemplo, *goma* en el idioma castellano (plástico) y en el español de Sudamérica (neumático).

La semántica estudia las diferentes relaciones existentes entre un signo y todos los demás dentro de un contexto llamado léxico, que es el que constituye un sistema que permite a los hablantes la interpretación, conocimiento, adquisición y uso de ese léxico.

- **Relaciones entre significantes: la homonimia**

Se dice que dos palabras son homónimas si su significante es el mismo; es decir, están compuestas por los mismos fonemas, o su realización fonética coincide. No se trata de relación entre significados. La relación homonímica más habitual se produce entre palabras de distinta categoría gramatical como se puede ver en la siguiente tabla.

| Homonimia (I) | |
|---|---|
| **Cojo** | Adjetivo. Persona o animal que cojea o le falta un pie o pierna. |
| **Cojo** | Primera persona del singular del presente indicativo del verbo coger. |

Tabla 1. Homonimia (I)

Pero también se produce en palabras de la misma categoría**.** Se da en aquellos casos en que el significado de las palabras no tiene ninguna relación, porque proceden de étimos distintos.

| Homonimia (II) | |
|---|---|
| **Granada** | Fruta |
| **Granada** | Proyectil |

Tabla 2. Homonimia (II)

Dentro del concepto general de homonimia, se pueden distinguir varios casos mostrados a continuación.

**Palabras homógrafas:** Tienen las mismas grafías y los mismos sonidos.

| Palabras homógrafas | |
|---|---|
| **Asta** | Cuerno |
| **Asta** | Palo de la bandera |

Tabla 3. Palabras homógrafas

**Palabras homófonas:** Tienen los mismos sonidos, pero distinta grafía.

| Palabras homófonas | |
|---|---|
| **Asta** | Palo de la bandera |
| **Hasta** | Preposición. |

Tabla 4. Palabras homófonas.

Todas ellas son homónimas. Las dos primeras son homógrafas, y las dos últimas son homófonas, entre sí, y respecto a las anteriores.

- **Separación de relaciones entre significado y significante: monosemia, polisemia y sinonimia**
  - **Monosemia**

Es la relación habitual que existe entre el significado y el significante en una palabra. A un significante se corresponde un solo significado.

Por ejemplo, la palabra *bolígrafo* expresa un referente que sólo puede ser evocado mediante ese significante.

- **Polisemia**

Una palabra es polisémica cuando se puede expresar con ella varios significados. O, dicho de otra forma: un significante puede tener varios significados.

La polisemia se distingue de la homonimia en que se trata de una relación entre los dos planos del signo lingüístico: los diferentes significados de una palabra tienen, o han tenido, un origen común.

| **Polisemia** | |
|---|---|
| **Vela** | Cilindro de cera o pieza de lona |
| **Cuarto** | Habitación o número fraccionario |

Tabla 5.Polisemia

La polisemia es uno de los mecanismos más eficaces de economía lingüística, pues permite expresar varios significados con un único significante.

- **Sinonimia**

Dos o más palabras son sinónimas si tienen el mismo significado. Es decir, la sinonimia consiste en la igualdad de significado, cuando existen diferentes significantes.

Se pueden distinguir diversas formas en que puede presentarse la sinonimia:

- **Sinonimia conceptual:** Los significados denotativos son plenamente coincidentes. Ej.: *mujer =esposa*
- **Sinonimia connotativa:** Puede, en ocasiones, no haber coincidencia denotativa; sin embargo, esto no impediría que se consideren sinónimos por los valores connotativos que encierran. Ej.: *espabilado =avispado.*
- **Sinonimia contextual:** En determinados contextos, se pueden establecer ciertas sinonimias que serían impensables en otros. Ej.: *pesado = indigesto*, el cocido es pesado.

➢ **Relaciones entre significados: el campo semántico**

- **Hiperonimia e hiponimia**

Se llama hiperónimo [7] a la palabra cuyo significado abarca al de otras, que se conocen como hipónimos [8]. Los hipónimos a los que se refiere una palabra son, entre sí, cohipónimos. Se pueden distinguir las relaciones siguientes ilustradas con ejemplos en tablas.

- **Relaciones de inclusión:** Un conjunto de palabras puede estar englobado dentro de otra palabra que las incluya a todas.

| Hiperónimo | Hipónimos | |
|---|---|---|
| Animal | Tigre | Hipónimos |
| | León | |
| | Lince | |

Tabla 6. Hipónimos

- **Relaciones lineales.** En otros casos, se establecen relaciones de sucesión. Así sucede, con los nombres de los meses o los días: *enero, febrero,..., diciembre; lunes, martes,..., domingo*.

| Hiperónimo | Hipónimos | |
|---|---|---|
| Meses | Enero | Cohipónimos |
| | Febrero | |
| | Marzo | |

Tabla 7. Cohipónimos

- **Valores expresivos del significado**

El significado puede convertirse en un elemento de máxima efectividad expresiva. Si se tienen en cuenta los elementos de la comunicación, la situación comunicativa aclarará el significado de muchas palabras. Pero a veces, el contexto referencial hará que surjan significados nuevos, que antes no estaban presentes.

Hay que tener en cuenta que toda palabra tiene un significado denotativo y un significado connotativo. Las connotaciones pueden ser positivas o negativas, siempre dependiendo del hablante que las considere.

Palabras tabú son aquellas que no se pronuncian, porque tienen una carga connotativa despectiva. Se sustituyen por otras palabras que designan la misma realidad, pero sin esas connotaciones peyorativas. Son los denominados eufemismos (del griego: palabra bien sonante).

En ciertas ocasiones de la historia, el uso de ciertas palabras puede herir la sensibilidad de la sociedad. Por ejemplo, en estos tiempos de crisis en los que nos encontramos, los medios de comunicación prefieren decir que una empresa ha hecho una *regulación de empleo*, a decir, un *despido masivo* para evitar el pánico colectivo.

Al igual que existen eufemismos, también se emplean disfemismos**.** Cuando la palabra tabú se sustituye por otra, pero de carácter humorístico. Ej: En lugar de *morir,* se utiliza el disfemismo *estirar la pata.*

**Pragmática**

La pragmática [9] es el estudio del modo en que el contexto influye en la interpretación del significado. El contexto debe entenderse como situación, ya que puede incluir cualquier aspecto extralingüístico.

Incluye en sus análisis los factores sociales, psicológicos, culturales, literarios, que determina la estructura de la comunicación verbal y sus consecuencias. En este nivel se relacionan la semántica y la sintaxis: la semántica hace abstracción de los usuarios y la sintaxis expresa la relación entre los signos sin tener en cuenta a los usuarios; sintetizando todo el proceso en el estudio del qué se dice y lo que literalmente se quiere decir.

Es fundamental analizar también las huellas que emisor y receptor dejan en el texto. Así, por ejemplo, la presencia de un YO que se dirige a un TÚ puede imprimir una cierta fuerza persuasoria al mensaje, al introducirse, consciente o inconscientemente, el autor en el texto en un intento de modificar la conducta de la persona que recibe el mensaje.

La pragmática busca analizar los cambios que se presentan en determinados contextos, porque estos contextos permean las palabras, los gestos y el mensaje en general de hablantes, y la o las interpretaciones hechas por los oyentes.

### 2.1.3 Problemas de la lingüística en el procesamiento del lenguaje natural

Los distintos niveles vistos anteriormente causan diversos problemas a los que se enfrenta el procesamiento del lenguaje natural. Algunos de los problemas son de difícil resolución y otros son irresolubles.

El problema que se da siempre es la ambigüedad, bien sea por su manera de escribirse (*banco,* entidad financiera; *banco,* asiento en el parque; *banco,* acumulación de arena en un rio) o bien por su manera de pronunciarse (*vaca*, animal; *baca*, sujeción del coche), como ya se vio en los apartados anteriores. La ambigüedad a su vez acarrea problemas que se van arrastrando a lo largo del análisis, tanto del morfológico como del sintáctico posteriormente. Con lo que es el principal problema al que se ha de enfrentar cualquier sistema de procesamiento de lenguaje natural.

En la siguiente tabla se muestra, a modo de resumen, los problemas que se generan por niveles de menor a mayor.

| Disciplina | Problema |
|---|---|
| **Fonología** | Afecta a los programas que utilicen reconocedor de voz Problemas con las pautas de voz de los usuarios Ej: Vaca / Baca Hora / Ora |
| **Morfología** | Confusiones a la hora de realizar el análisis morfológico en palabras que se escriben igual (homógrafas) Ej: "La casa" →verbo casar<br>→ edificio |
| **Sintaxis** | Concordancia de género y número |
| **Semántica** | La homonimia y la polisemia |
| **Pragmática** | Desconocimiento del contexto Problemas con la persona.<br>Ej: "¿Tienes hora?" |

Tabla 8. Problemas de cada disciplina

## 2.2 Procesamiento del Lenguaje Natural (PLN)

### 2.2.1 Introducción

El procesamiento del lenguaje natural o Natural Language Processing(PLN o NLP) [10] es una disciplina de la ingeniería lingüística que se encarga de estudiar de qué manera se puede mejorar el lenguaje natural con un lenguaje que entienda un ordenador, a través de algún dispositivo que interaccione con el usuario ya sea mediante texto o voz.

Por lo tanto el objetivo del PLN es procesar, traducir e interpretar de forma automática el lenguaje de los humanos para que, a través de una serie de algoritmos, las máquinas sean capaces de extraer la información correctamente. Es decir, es hacer una serie de sistemas y programas que sean capaces de recuperar la información que un usuario humano introduce en su propio idioma, (inglés, francés, español, etc.).

El procesamiento del lenguaje natural presenta múltiples aplicaciones:

- Corrección de textos
- Traducción automática
- Recuperación de la información
- Extracción de información y resúmenes
- Búsqueda de documentos
- Sistemas Inteligentes para la educación y el entrenamiento

La **corrección de textos** permite la detección y corrección de errores ortográficos y gramaticales. Para detectar este tipo de errores, la computadora necesita entender en cierto grado el sentido del texto. Los correctores de gramática detectan las estructuras incorrectas en las oraciones aunque todas las palabras en la oración estén bien escritas en el lenguaje en cuestión. El problema de detectar los errores de este tipo es complejo debido a la existencia de gran variedad de estructuras permitidas.

Para describir las estructuras de las oraciones en el idioma, se usan las llamadas gramáticas formales, o sea conjuntos de reglas de combinación de palabras y su orden relativo en las oraciones.

La **traducción automática** se refiere a la traducción correcta de un lenguaje a otro, tomando en cuenta lo que se quiere expresar en cada oración.

En el campo de la **recuperación de la información** se han desarrollado sistemas que permiten obtener información sobre estadísticas deportivas, información turística, geografía etc. En lugar de buscar los documentos para encontrar en ellos la respuesta a su pregunta, el usuario podría hacer su pregunta a la computadora: *¿Cómo se llama el presidente de Francia?, ¿Cuáles son los centros más avanzados en Procesamiento del Lenguaje Natural?,* y otras.

Por otra parte, se han desarrollado sistemas con la capacidad de **crear resúmenes** de documentos a partir de los datos suministrados. Estos sistemas son capaces de realizar un análisis detallado del contenido del texto y elaborar un resumen.

También se han desarrollado sistemas inteligentes que **permiten modelar el comportamiento** del estudiante, reconocer y procesar sus errores, desarrollar habilidades en la resolución de problemas y otras actividades del proceso enseñanza y aprendizaje. En ellos el Procesamiento del Lenguaje Natural juega un papel de relevante importancia en la creación y desarrollo de interfaces amigables**.**

### 2.2.2 Arquitectura típica de un sistema de Procesamiento del Lenguaje Natural

Uno de los elementos fundamentales en el diseño de un sistema PLN es sin lugar a dudas la determinación de la arquitectura del sistema, es decir, como se introducen los datos a la computadora y como ella interpreta y analiza las oraciones que le sean proporcionadas. A continuación se muestra un esquema del análisis léxico/sintáctico por computadora. El sistema consiste de:

- El usuario le expresa (de alguna forma) a la computadora que tipo de procesamiento desea hacer.
- La computadora analiza las oraciones proporcionadas, en el sentido morfológico y sintáctico;
- Luego, se analizan las oraciones semánticamente, es decir se determina el significado de cada oración;
- Se realiza el análisis pragmático del texto. Así, se obtiene una expresión final.

Se ejecuta la expresión final y se entrega al usuario para su consideración.

A continuación se muestra de una forma gráfica los pasos que se realizan de forma típica a la hora de implementar un PLN.

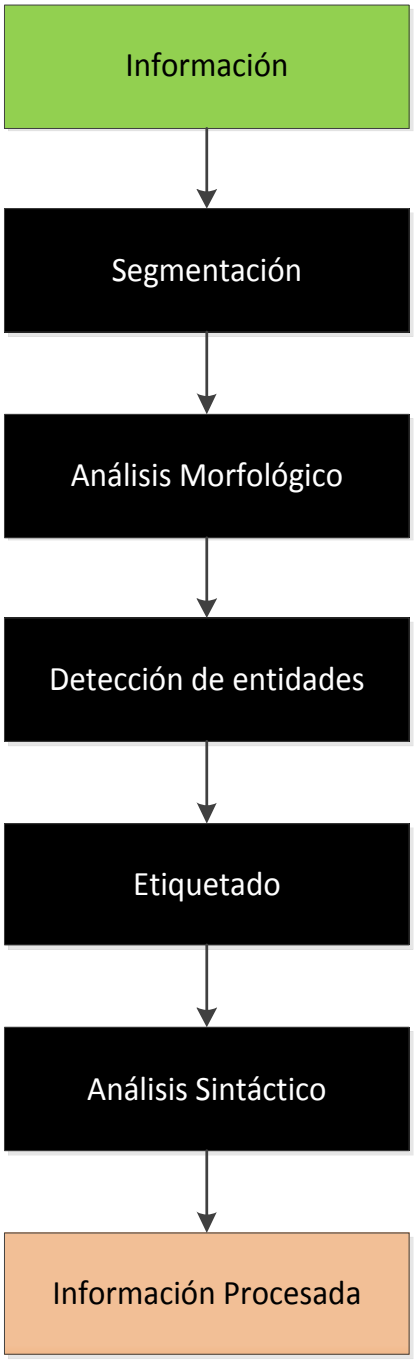


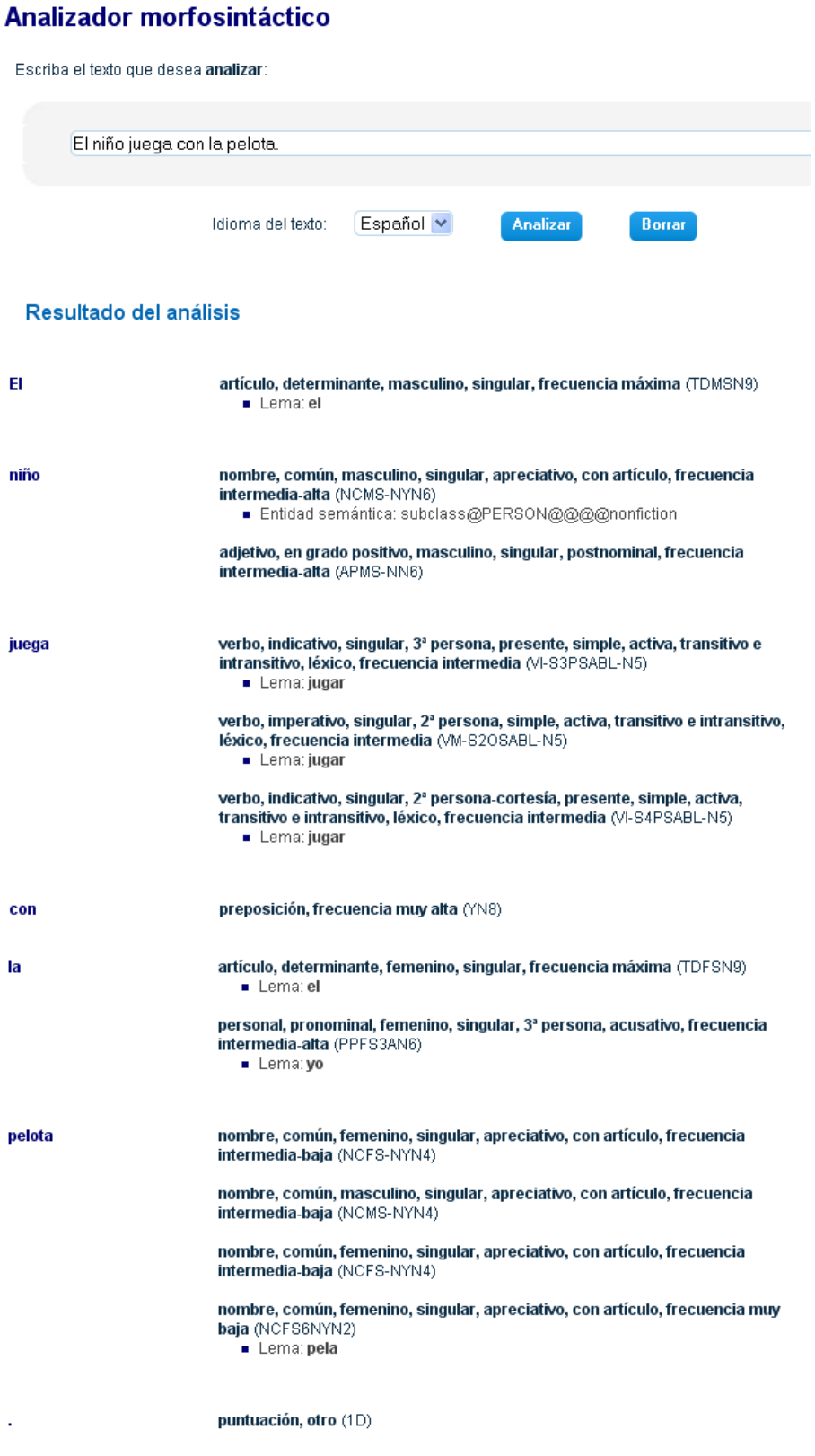

Imagen 1. Esquema de procesado típico PNL

### 2.2.3 Segmentación

Principalmente, la segmentación consiste en dividir el texto por frases, y estas frases a su vez en palabras. De este modo, una vez se tienen las palabras por separado, puedan ser categorizadas y se interprete el significado que aportan al conjunto del que son parte.

### 2.2.4 Análisis Morfológico

El análisis morfológico consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración. Un ejemplo es el mostrado en la figura siguiente (utilizando la herramienta STILUS [11]).



## Analizador morfosintáctico

Escriba el texto que desea **analizar**:

El niño juega con la pelota.

Idioma del texto: Español Analizar Borrar

### Resultado del análisis

**El** — **artículo, determinante, masculino, singular, frecuencia máxima** (TDMSN9)
- Lema: **el**

**niño** — **nombre, común, masculino, singular, apreciativo, con artículo, frecuencia intermedia-alta** (NCMS-NYN6)
- Entidad semántica: subclass@PERSON@@@@nonfiction

**adjetivo, en grado positivo, masculino, singular, postnominal, frecuencia intermedia-alta** (APMS-NN6)

**juega** — **verbo, indicativo, singular, 3ª persona, presente, simple, activa, transitivo e intransitivo, léxico, frecuencia intermedia** (VI-S3PSABL-N5)
- Lema: **jugar**

**verbo, imperativo, singular, 2ª persona, simple, activa, transitivo e intransitivo, léxico, frecuencia intermedia** (VM-S2OSABL-N5)
- Lema: **jugar**

**verbo, indicativo, singular, 2ª persona-cortesía, presente, simple, activa, transitivo e intransitivo, léxico, frecuencia intermedia** (VI-S4PSABL-N5)
- Lema: **jugar**

**con** — **preposición, frecuencia muy alta** (YN8)

**la** — **artículo, determinante, femenino, singular, frecuencia máxima** (TDFSN9)
- Lema: **el**

**personal, pronominal, femenino, singular, 3ª persona, acusativo, frecuencia intermedia-alta** (PPFS3AN6)
- Lema: **yo**

**pelota** — **nombre, común, femenino, singular, apreciativo, con artículo, frecuencia intermedia-baja** (NCFS-NYN4)

**nombre, común, masculino, singular, apreciativo, con artículo, frecuencia intermedia-baja** (NCMS-NYN4)

**nombre, común, femenino, singular, apreciativo, con artículo, frecuencia intermedia-baja** (NCFS-NYN4)

**nombre, común, femenino, singular, apreciativo, con artículo, frecuencia muy baja** (NCFS6NYN2)
- Lema: **pela**

**.** — **puntuación, otro** (1D)

Imagen 2. Ejemplo de análisis morfosintáctico con la herramienta STILUS

Como se aprecia en la ilustración, realizar el análisis morfológico realmente consiste en dar una etiqueta o marca a cada palabra, bien sea un código que interpretará posteriormente el programa o bien sean dándole los rasgos completos.

A la hora de realizar este etiquetado, surge un problema que después jugará un papel muy importante para realizar correctamente los análisis o no, la desambigüación.

En el caso de STILUS, el etiquetado consiste en asignar una marca (etiqueta) formada por cifras y letras en las que se indica la categoría gramatical, género, número, persona, tiempo verbal, modo…

Esto ayuda a elegir las palabras que se van a buscar ya que es muy común dar más importancia a un nombre que a una preposición.

Para el caso particular de este proyecto se optó por la opción de dar un código a cada palabra, más adelante, en el apartado del diseño se especificará que tipo de palabra se corresponde con su código.

### 2.2.5 Detección de Entidades

Algunos sistemas y herramientas del procesamiento del lenguaje natural son capaces de detectar entidades tipo fechas, ciudades, personas, etc. Estas detecciones permiten encontrar multipalabras que se refieren a un mismo objeto pero que se encuentra fraccionado, como las fechas como se aprecia en el ejemplo, o los nombres propios. Ejemplo: "*Rafael Nadal*" se debe referir a la misma persona, no "*Rafael*" por una parte al cantante y "*Nadal*" al tenista o jugador de fútbol.

En la siguiente imagen se muestra un ejemplo de detección de entidades que hace la herramienta STILUS.

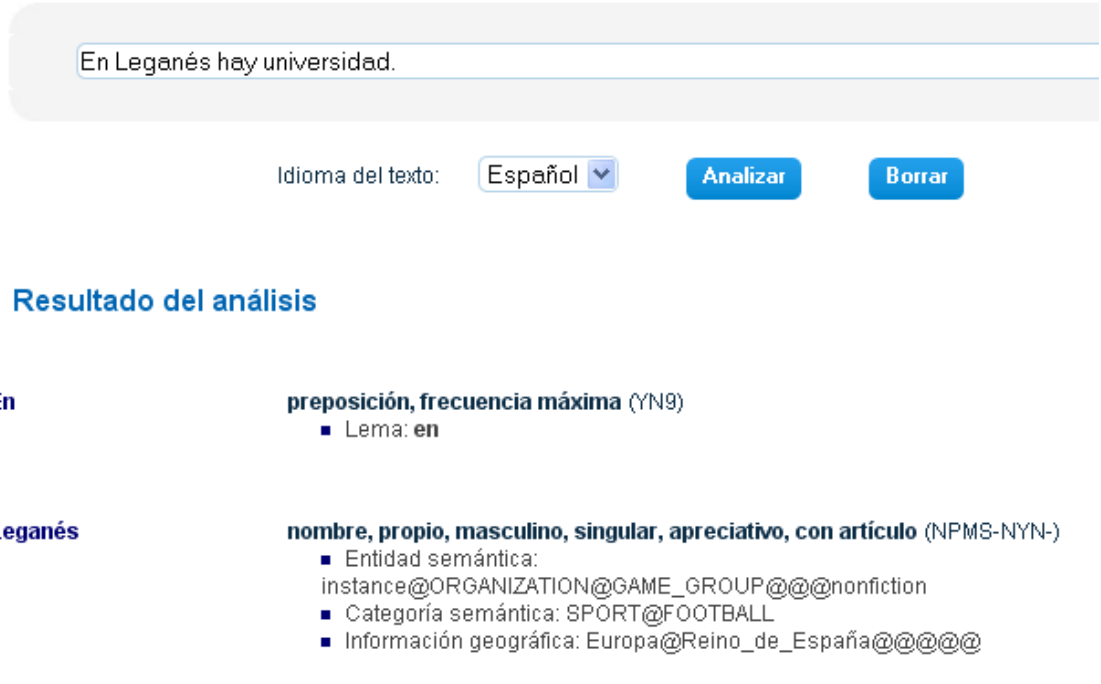



Imagen 3. Ejemplo de detección de entidades con la herramienta STILUS.

En este proyecto se usa la detección de entidades de tipo nombres propios, los cuales deberán estar escritos correctamente. En el apartado de diseño se detallarán las principales características de detección de entidades desarrolladas en este proyecto.

### 2.2.6 Etiquetado

Este etiquetado se basa en lematizar cada una de las palabras en las que se ha fraccionado el texto. Lematizar consiste en la reducción de las diferentes formas flexivas de una palabra a la forma canónica, su lema o representante de toda la familia flexiva.

Por convenio esta reducción consiste en reagrupar las distintas inflexiones de un verbo en el infinitivo; el singular y el plural de un sustantivo en el singular; el masculino y el femenino de un adjetivo en el masculino. Con esto se consigue identificar familias de palabras para considerarlas como una sola. Es un dato muy importante para la búsqueda de información ya que se consigue hacer independiente la búsqueda de tiempos verbales y otros.

### 2.2.7 Análisis Sintáctico

De todos los niveles de análisis, la sintaxis ha sido durante mucho tiempo y aún sigue siendo el nivel al que la lingüística le ha prestado mayor atención. Esta casi exclusiva atención se justifica por dos razones principales en cuanto al tratamiento automático del lenguaje natural:

- El procesamiento semántico funciona sobre los constituyentes de la oración. Si no existe un paso de análisis sintáctico, el sistema semántico debe identificar sus propios constituyentes. Por otro lado, si se realiza un análisis sintáctico, se restringe enormemente el número de constituyentes a considerar por el semántico, mucho más complejo y menos fiable. El análisis sintáctico es mucho menos costoso computacionalmente hablando que el análisis semántico (que requiere inferencias importantes). Por tanto, la existencia de un análisis sintáctico conlleva un considerable ahorro de recursos y una disminución de la complejidad del sistema.
- Aunque frecuentemente se puede extraer el significado de una oración sin usar hechos gramaticales, no siempre es posible hacerlo. La sintaxis contempla dos modos diferentes, pero no por ello opuestos, de análisis. El primero es el análisis

de constituyentes o análisis de estructura de frase: la estructuración de las oraciones en sus partes constituyentes y la categorización de estas partes como nominales, verbales, adjetivales, etc. El segundo es el análisis de las relaciones o funciones gramaticales: la asignación de relacionales gramaticales tales como sujeto, objeto, etc.

### 2.2.8 Análisis Semántico

En muchas aplicaciones del PLN los objetivos del análisis apuntan hacia el procesamiento del significado. En los últimos años las técnicas de procesamiento sintáctico han experimentado avances significativos, resolviendo los problemas fundamentales, sin embargo, las técnicas de representación del significado no han obtenido los resultados deseados, y numerosas cuestiones continúan sin encontrar soluciones satisfactorias.

Definir qué es el significado no es una tarea sencilla, y puede dar lugar a diversas interpretaciones. A efectos prácticos es necesaria una buena modularidad para facilitar el procesamiento, de tal manera que sea posible distinguir entre significado independiente o dependiente del contexto. El significado independiente del contexto, tratado por la semántica, hace referencia al significado que las palabras tienen por sí mismas sin considerar el significado adquirido según las circunstancias en las que se está usando dicha palabra, es decir, se hace referencia a las condiciones de verdad de la frase, ignorando la influencia del contexto o las intenciones del hablante. Por su parte, el significado dependiente del contexto, estudiado por la pragmática, se refiere al componente significativo de una frase asociado a las circunstancias en que ésta se utiliza.

Atendiendo al desarrollo en el proceso de interpretación semántica, es posible optar entre múltiples pautas para su organización, explicadas a continuación.

En referencia a la estructura semántica que se va a generar, puede interesarnos que exista una simetría respecto a la estructura sintáctica, de tal manera que se generará una estructura arbórea para el análisis semántico que tendrá las mismas características que el árbol sintáctico, o por el contrario que no se dé tal correspondencia entre ellas, caso en el que se realizarán varias transformaciones sobre la estructura utilizada en la sintaxis, generándose la representación semántica sobre dichas transformaciones.

Cada una de las dos opciones anteriores puede implementarse de forma secuencial (en primer lugar se realiza el análisis sintáctico y, una vez finalizado éste, se pasa al análisis semántico) o paralela (se puede iniciar el análisis semántico de cada constituyente cuando éste ha sido tratado por el analizador sintáctico).

Finalmente, en combinación con cada una de las opciones anteriores, podemos escoger un modelo en el que exista una correspondencia entre reglas sintácticas y semánticas o, contrariamente, podemos optar por un modelo que no cumpla tal requisito. En caso afirmativo, para cada regla sintáctica existirá una regla semántica correspondiente.

El significado es representado por formalismos conocidos por el nombre de *knowledge representation*. El léxico proporciona el componente semántico de cada palabra en un formalismo concreto, y el analizador semántico lo procesa para obtener una representación del significado de la frase.

## 2.3 XML

Las siglas en inglés XML[12] provienen de "*e**X**tensible **M**arkup **L**anguage*" lo que traducido al español es un lenguaje de marcas extensible. XML fue creado por el *World Wide Wev Consortium*[13] (W3C) como un metalenguaje extensible de etiquetas.

Es una simplificación y adaptación del SGML y permite definir la gramática de lenguajes específicos. De esta forma XML no es en realidad un lenguaje específico, sino una forma de definir lenguajes para distintas necesidades, es por ello que se le llama metalenguaje. Algunos de estos lenguajes son XHTML, SVG, por ejemplo.

Se podría considerar que XML es un lenguaje de metamarcado que ofrece un formato para la descripción de datos estructurados. Esto facilita unas declaraciones de contenido más precisas y unos resultados de búsquedas más significativos en varias plataformas, con esto se propone como un estándar para el intercambio de información estructurada. XML se puede usar en editores de texto, bases de datos, hojas de cálculo…

XML es una tecnología sencilla que tiene a su alrededor otras que la complementan y la hacen mucho más grande y con unas posibilidades mucho mayores. Tiene un papel muy importante

en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

### 2.3.1 Ventajas de XML

XML presenta múltiples ventajas, a continuación se muestran las más relevantes:

- Es extensible. Después de diseñado y puesto en producción, es posible extender XML con la adición de nuevas etiquetas, de modo que se pueda continuar utilizando sin complicación alguna.
- El analizador es un componente estándar, no es necesario crear un analizador específico para cada versión de lenguaje XML. Esto posibilita el empleo de cualquiera de los analizadores disponibles. De esta manera se evitan *bugs* y se acelera el desarrollo de aplicaciones.
- Si un tercero decide usar un documento creado en XML, es sencillo entender su estructura y procesarla. Por lo tanto se mejora la compatibilidad entre aplicaciones. Se pueden comunicar aplicaciones de distintas plataformas, sin que importe el origen de los datos.
- Existe una transformación de datos en información, pues se le añade un significado concreto y los asociamos a un contexto, con lo cual tenemos flexibilidad para estructurar documentos.
- Los autores y proveedores pueden diseñar sus propios tipos de documentos usando XML.
- La información contenida puede ser más 'rica' y fácil de usar, porque las habilidades hipertextuales de XML son amplias.
- XML puede dar más y mejores facilidades para la representación en los visualizadores.
- La información es más accesible y reutilizable, porque la flexibilidad de las etiquetas de XML pueden utilizarse sin tener que amoldarse a reglas específicas.
- Los archivos XML válidos son válidos también en SGML, luego pueden utilizarse también fuera de la Web, en un entorno SGML (una vez la especificación sea estable y el software SGML la adopte).
- Elimina muchas de las complejidades de SGML, en favor de la flexibilidad del modelo, con lo que la escritura de programas para manejar XML será más sencilla que haciendo el mismo trabajo en SGML.

- Es muy sencillo de usar, además cuenta con múltiples tutoriales en Internet, como puede ser el “w3cshools” [14].

### 2.3.2 Estructura de un documento XML

Aunque a primera vista, un documento XML puede parecer similar a HTML, hay una diferencia principal. Un documento XML contiene datos que se autodefinen, exclusivamente. Un documento HTML contiene datos mal definidos, mezclados con elementos de formato. En XML se separa el contenido de la presentación de forma total.

Una forma de entender rápidamente la estructura de un documento XML es viendo un pequeño ejemplo:

```
<?xml version="1.0"?>
<!DOCTYPE MENSAJE SYSTEM "mensaje.dtd">
<mensaje>
<remite>
<nombre>Alfredo Reino</nombre>
<email>alf@ibium.com</email>
</remite>
<destinatario>
<nombre>Bill Clinton</nombre>
<email>president@whitehouse.gov</email>
</destinatario>
<asunto>Hola Bill</asunto>
<texto>
<parrafo>¿Hola qué tal? Hace <enfasis>mucho</enfasis> que no escribes. A ver si llamas y quedamos para tomar algo.</parrafo>
</texto>
</mensaje>
```

Este mismo documento puede ser visto de forma gráfica, para comprender mejor la estructura de un documento XML.

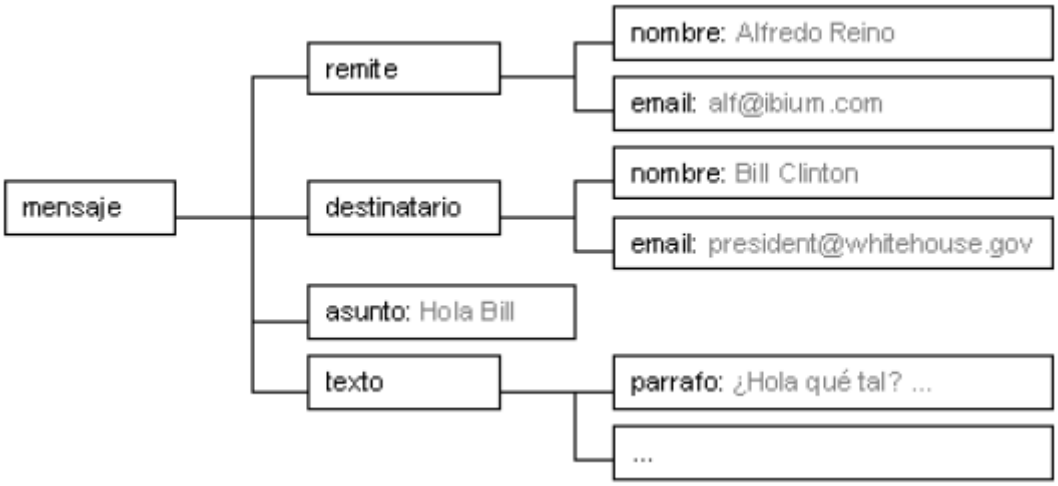



Imagen 4. Ejemplo árbol XML.