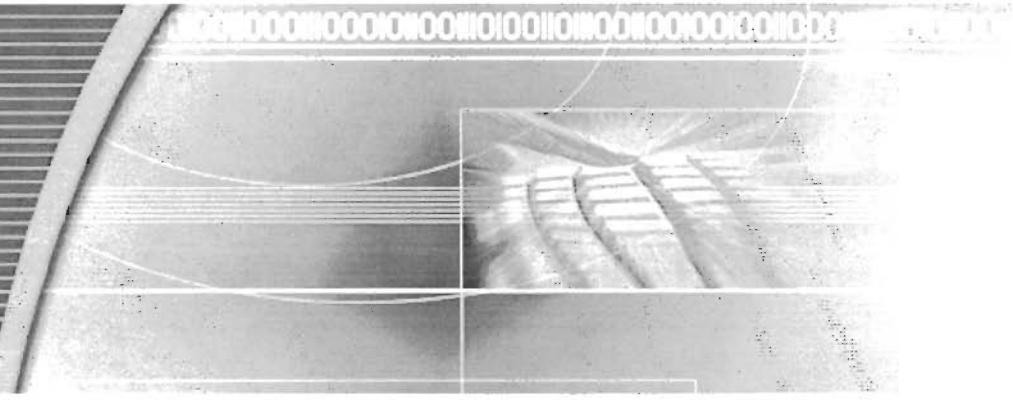
MATCOM

Facultad de Matemática - Computación

Sistema de Consulta en Lenguaje Natural a Ficheros de Preguntas Frecuentes



Trabajo de Diploma

optando por el título de Licenciado en Ciencia de la Computación

Autor: **Ediber Montoya Moreira**

Tutor: **Ing. William Azcuy Morales**

UCI

Informáticas

Ciudad de la Habana, julio del 2004

Resumen

En este informe se realiza un estudio de los *Sistemas de Consultas en Lenguaje Natural a Ficheros de Preguntas Frecuentes* y se profundiza en los aspectos lingüísticos y computacionales en el tratamiento de la morfología para la construcción de un Traductor Léxico.

Los sistemas con estas características son una respuesta a las necesidades de los usuarios de consultar los extensos ficheros de preguntas frecuentes (FAQ) asociados a sitios publicados en Internet y a aplicaciones informáticas. Y sirven de interfaz en lenguaje natural a una base de conocimientos cuyo contenido son pares de preguntas-respuestas.

Con este trabajo se obtiene un lenguaje y un prototipo de compilador para la construcción de un Traductor Léxico cuya representación léxica se realiza mediante *clases de palabras* que agrupan palabras con similar comportamiento morfotáctico e igual caracterización morfosintáctica. Otro resultado es la ampliación de las reglas de dos niveles con contextos morfosintácticos expresados en términos de rasgos morfosintácticos y clases de palabras. En el caso de la morfotáctica proponemos su especificación mediante una gramática lineal a la derecha y sin ciclos, con el objetivo de mantener la posibilidad de conversión en un Traductor de Estados Finitos y resolver el problema de la dependencia a distancia.

Introducción

Una persona sentada frente a una computadora espera soluciones rápidas a sus problemas, igualmente una persona frente a un sitio WEB o una aplicación informática espera encontrar respuestas rápidas a sus preguntas.

Habitualmente, las preguntas más usuales que se podría hacer un usuario se encuentran publicadas y, para encontrarlas, es necesario recorrer un extenso fichero con preguntas que pueden, incluso, encontrarse sin un orden lógico o, en su defecto, demasiado reducidos como para contener la pregunta buscada. Los Sistemas de Consultas en Lenguaje Natural a Ficheros de Preguntas Frecuentes (FAQ-Systems) vienen a ser una solución a esta problemática aportando una interfaz cómoda que permita localizar la pregunta en cuestión, o una equivalente, a partir de su texto.

A primera vista esta tarea, aparentemente trivial, podría pensarse que se limita sólo a hacer una búsqueda en el fichero del texto correspondiente a la pregunta; pero un asomo a la riqueza semántica de cualquier lenguaje natural muestra que un simple "matching" entre cadenas de caracteres no es suficiente.

Podría optarse por emplear un Sistema de Búsqueda de Respuesta tradicional que utiliza la WEB como base de conocimientos y busca, dentro de las páginas devueltas por un motor de búsqueda, los segmentos de texto que pudieran considerarse respuesta a la pregunta de entrada. Varios trabajos se han realizado en esta área: [Mollá et. al 1998], [Vicedo & Fernández 2000], [Burger et. al 2001], [Light et. al 2001], [Tou et. al 2001]. De forma general, estos sistemas no obtienen buenos resultados, sólo aquellos que realizan el análisis sobre un dominio restringido han logrado respuestas más o menos aceptables en términos generales.

En la actualidad, el problema de buscar una respuesta dentro de un fichero de preguntas frecuentes no está del todo resuelto aunque se han desarrollado prototipos como: [Burke et. al. 1997] y [Sneiders], que han demostrado su eficacia. Estos dos sistemas, que se realizaron para el inglés son los que se toman como base para el estudio inicial en el desarrollo del presente trabajo aprovechando que sus modelos no dependen de una lengua en específico. Los enfoques que emplean son bastante distintos entre ellos. En el primero de los casos se opta por combinar métodos estadísticos y procesamiento de lenguaje natural a un nivel superficial apoyado con un módulo que determina la cercanía entre dos palabras buscando el punto de coincidencia más cercano en sus árboles de hiperónimos. El segundo, opta por representar toda la información lingüística con el par pregunta-respuesta, incluyendo un pequeño diccionario con todas las posibles flexiones de los términos de la pregunta. En este diccionario se incluyen, además, los posibles sinónimos restringidos al contexto impuesto por la pregunta. La tarea de comparar las preguntas se realiza mediante el concepto: Prioritized Keyword Matching.

De modo general, en el trabajo que desarrollamos se persigue el siguiente objetivo:

- Construir un Sistema que busque dentro de un fichero de pares pregunta-respuesta las preguntas más parecidas a una pregunta de entrada y retorne sus respuestas.

Para esta etapa los objetivos específicos son los siguientes:

- Realizar un estudio teórico sobre los Sistemas de Consulta en Lenguaje Natural a Ficheros de Preguntas Frecuentes.
- Realizar un estudio profundo de la morfología del español, resaltando sus aspectos esenciales.
- Realizar un estudio de los acercamientos computacionales a la morfología.
- Proponer un lenguaje para la representación del léxico del español.
- Proponer o utilizar un formalismo para construir un Analizador Morfológico.
- Construir un módulo de prueba para validar el formalismo utilizado.

Cómo se puede apreciar, los objetivos específicos se limitan a enunciar los lineamientos esenciales para la construcción del sistema final y profundizar en el estudio de la morfología como base para superiores niveles de análisis desde el punto de vista del procesamiento de lenguaje natural.

La morfología ha encontrado en la propuesta de Kimmo Koskenniemi, enunciada a principio de los años 80': "Morfología de dos niveles", un formalismo potente y capaz de expresar de forma clara los procesos de formación de palabras en las lenguas con morfología concatenativa. No obstante su éxito, a esta propuesta inicial, se han incorporado varias modificaciones dirigidas a mejorar su eficiencia y expresividad. Uno de las mejoras indiscutibles a este modelo son los Traductores Léxicos, en los cuales se componen junto con el lexicón las reglas de dos niveles. En este tema se propone un lenguaje que permite expresar de forma clara las diferentes clases de palabras desde el punto de vista morfotáctico, asociándoles rasgos morfosintácticos. Para la representación de la morfotáctica se emplea una gramática lineal a la derecha y sin ciclos que permite resolver los problemas de la dependencia a distancia y extender el tradicional mecanismo de clases de continuación. A partir de los planteamientos expuestos en [Badía et. al], las reglas de dos niveles son ampliadas con rasgos morsintácticos, permitiendo además asociarlas a las clases de palabras representadas en el léxico y, para hacer más claras las restricciones, se permite su aplicación en ciertos puntos en la gramática.

El informe está estructurado en cuatro capítulos:

- **Capítulo 1. Fundamentación Teórica:** En este capítulo se realiza un estudio del estado del arte a partir de los trabajos mencionados anteriormente. Se realiza un análisis del problema desde la teoría de conjunto y finalmente se describen de modo superficial las herramientas computacionales necesarias para el procesamiento de lenguaje natural que llevaría el sistema.
- **Capítulo 2. Consideraciones sobre la Morfología de la lengua española:** En este capítulo se describen las principales características de los procesos de formación de palabras, se describen los modelos lingüísticos empleados

tradicionalmente para modelar la morfología y finalmente se describen las principales clases de palabras desde el punto de vista morfosintáctico.

- **Capítulo 3. La morfología desde el punto de vista computacional:** Este capítulo describe diferentes acercamientos a la morfología y profundiza en el formalismo conocido como "Morfología de dos Niveles".

- **Capítulo 4. La Construcción del Traductor Léxico:** En este capítulo se describen los cambios propuestos al formalismo de dos niveles y se realiza una descripción del lenguaje empleado para la representación léxica y los detalles del proceso de compilación del Traductor Léxico.

Capítulo 1.

Fundamentación Teórica

Introducción

En este capítulo realizaremos un estudio de los sistemas: *FAQ Finder* [Burke et. Al 1997] y *Automated FAQ Answering* [Sneiders] que tienen características similares al proyecto que desarrollamos y, además, realizamos un análisis del problema a partir de la teoría de conjuntos en función de qué información permitiría describir una pregunta y cómo representarla. Adicionalmente se describen de forma superficial las herramientas computacionales necesarias para determinar el grado de parecido entre dos preguntas en lenguaje natural.

1.1 Visión del Problema desde la Teoría de Conjuntos

Desde el punto de vista de la teoría de conjuntos elo problema consiste en dado una pregunta en el idioma español encontrar su respuesta a partir de una Base de Conocimientos (BC) compuesta por pares (x,y) donde "x" es una oración interrogativa en el idioma español y "y" es un conjunto ordenado de oraciones del mismo idioma que desde el punto de vista semántico se considera respuesta a la pregunta "x". Ahora, para determinar cuál es la respuesta correcta a la pregunta hecha, de cada par sólo se considera la primera parte. Luego el problema se reduce a determinar dada una pregunta de entrada, cuál es la pregunta almacenada en la BC que le es semánticamente equivalente y retornar la respuesta correspondiente.

Veamos algunas consideraciones relacionadas con el problema desde el punto de vista de la teoría de conjuntos.

Sea L el conjunto de oraciones del idioma español.

Sea Q, subconjunto de L, el conjunto de las oraciones interrogativas de L.

El problema que nos ocupa es determinar dado (x, y) ambas pertenecientes a Q si (x R y) donde R se define de la siguiente forma: (x R y) si "x" es semánticamente equivalente a "y".

Si asumimos como cierto, al menos preliminarmente, el hecho de que "*R*" es una relación de equivalencia definida sobre Q, entonces Q se puede particionar en clases de equivalencia y el problema se resumiría a determinar dada una pregunta cuál es su clase de equivalencia. Desde este punto de vista surgen dos algoritmos generales:

a) Asumiendo la arquitectura inicial de la BC.

Determinar la clase de equivalencia de la pregunta de entrada

Por cada pregunta en la BC determinar su clase de equivalencia y si corresponde con la clase de la pregunta de entrada entonces retornar su respuesta.

b) Este algoritmo implica cambios en la arquitectura de la BC, pues asume que en ella se almacena en vez de la pregunta con su respuesta, la clase de equivalencia de la pregunta, con la respuesta.

Determinar la clase de equivalencia de la pregunta de entrada.

Buscar la clase de equivalencia en la BC y retornar su respuesta.

En estos algoritmos el punto crítico reside en determinar la clase de equivalencia de una pregunta, y en el "b)" la particularidad está en que se pierde información semántica al no representar la pregunta original asociada a la respuesta, aunque elimina tener que buscar, en ese momento, su clase de equivalencia. Una solución sería considerar la BC como una

terna (x, y, z) donde "x" sea la pregunta original, "y" la clase de equivalencia a la que pertenece y "z" la respuesta correspondiente.

Concentrémonos entonces en determinar dado una pregunta cuál es su clase de equivalencia.

Una clase de equivalencia es un conjunto, en este caso subconjunto propio de Q y para definirlos se puede hacer de forma extencional o intencional. Su representación extencional implicaría contar con una lista de todas las preguntas que forman parte de ella y en la práctica sería una tarea bastante engorrosa.

Ej.

Para la oración:

¿Cómo instalar Windows en mi computadora?

Podríamos considerar equivalentes las siguientes oraciones:

¿Cómo puedo instalar Windows en mi computadora?

¿Cómo implantar Windows en mi computadora?

¿Cómo puedo implantar Windows en mi computadora?

¿Cómo instalar el Sistema Operativo Windows en mi computadora?

Y como estas otras muchas que se diferenciarían unas de otras en los sinónimos que empleen, en la introducción de información redundante o en el orden de los términos.

Por otra parte, la representación intencional implicaría contar con un generador o el conjunto mínimo de características que lo identifican unívocamente, estas características podrían ser sus constituyentes desde el punto de vista sintáctico, sustituyendo cada termino por su clase de equivalencia semántica (sinónimos).

Ej.

Tomemos la misma oración vista en el ejemplo anterior:

¿Cómo instalar Windows en mi computadora?

Una posible descripción de esta oración sería la siguiente:

Tipo de pregunta: *Cómo*

Núcleo del Sintagma Verbal: *instalar*

Núcleo del Sintagma Nominal: *Windows*

Si a este núcleo le adicionamos algunos complementos y también permitimos sustituir los núcleos de los sintagmas obtendríamos una gran lista de oraciones con un significado *bastante parecido* al de la oración original.

Entre nuestras premisas estaba el hecho de considerar la relación "*R*" como una relación de equivalencia y, en la práctica, esta premisa no se cumple porque adicionalmente a las ambigüedades morfosintácticas y semánticas inherentes a la oración, siempre está presente la ambigüedad relativa al discurso que tiene que ver con el contexto en que se expresa la pregunta [Potier 1970]. De esto resulta que en la práctica "*R*" no es una relación de equivalencia, por lo que el modelo planteado no es viable. Ahora bien, si relajamos "*R*" considerándola, ya no como una equivalencia sino, distancia o grado de parecido entre dos preguntas conseguimos, no una partición, pero sí, una segmentación del conjunto "*Q*" en subconjuntos, naturalmente no disjuntos, pero donde el cardinal de la intercepción sería mínimo o cercano a este, planteando el cardinal mínimo como una representación numérica de la ambigüedad. Entonces el problema se convierte en determinar la ecuación que define a "*R*" y definir un umbral que marque la exclusión de las dos preguntas.

Para definir a *"R"* parece viable utilizar los datos aportados por la representación intencional del conjunto al que pertenece. No obstante, no daremos una opinión concluyente al respecto principalmente por no contar con todo el conocimiento lingüístico necesario para hacerlo.

1.2 Estudio del Estado del Arte

En la investigación preliminar se encontraron dos sistemas con características semejantes al proyecto que desarrollamos, ambos para el inglés, son estos: *FAQ Finder* [Burke et. Al 1997] y *Automated FAQ Answering* [Sneiders], cuyas características se describen a continuación.

1.2.1 FAQ Finder

FAQ Finder [Burke et. al 1997] es un sistema implementado en Lisp que recibe preguntas expresadas en Lenguaje Natural y retorna una respuesta utilizando como base de conocimientos ficheros compuestos por pares preguntas-respuestas. Este sistema se implementa utilizando una combinación de técnicas estadísticas de RI, la base de conocimientos semánticos WordNet y procesamiento del Lenguaje Natural a un nivel superficial.

Sus características principales son las siguientes:

-Utiliza ficheros que contienen una secuencia de pares Pregunta-Respuesta (P&R).

-Toda la información útil para determinar la relevancia de un par P&R se encuentra con el par.

-De los pares P&R la parte correspondiente a la pregunta es la más relevante para determinar el grado de parecido con la pregunta del usuario.

-Sólo necesita un conocimiento parcial del lenguaje para el macheo entre las preguntas.

La base de conocimientos WordNet no se emplea en su totalidad, sino que se limita sólo al uso de las relaciones de sinonimia e hiperonimia a partir de las cuales se construye un árbol asociado a cada término y el grado de parecido se obtiene buscando el punto de coincidencia menos profundo entre los árboles.

Este sistema está diseñado para procesar automáticamente las páginas de Preguntas Frecuentes que se pueden encontrar en muchos sitios en INTERNET.

1.2.2 Automated FAQ Answering.

El sistema Automated FAQ Answering [Sneiders] enfoca la tarea de comparación entre preguntas a partir del concepto: Prioritized Keyword Matching. El cual, básicamente, consiste en mantener junto al par Pregunta-Respuesta una estructura que contiene tres conjuntos de palabras:

- required words
- optional keywords
- forbidden keywords

En el primer conjunto: *required words* se especifican las palabras que deben aparecer de forma obligatoria en la pregunta, la ausencia de al menos una de ellas provoca que las preguntas no se consideren semejantes. En el tercer conjunto: *forbidden keywords*, se incluyen las palabras que no pueden aparecer en la pregunta y de existir al menos una de ellas se rechaza la pregunta como semejante. El tratamiento de las *optionals keywords* es un poco más complejo. En el tratamiento de la pregunta de entrada se excluyen las llamadas "*stop words*", en esta categoría aparecen las palabras que carecen de gran importancia estadística, como los artículos, conjunciones, etc. Para las restantes, las que no aparecen entre las opcionales y no se incluyen en la lista de *stop words* se establece un límite específico para cada pregunta. Sobrepasar ese límite da como resultado el rechazo de la pregunta como semejante.

Este sistema plantea también el tratamiento del Lexicón a nivel local, o sea, individual para cada pregunta almacenada, y su representación se realiza en forma de léxico desplegado (lista de formas y sinónimos). Por cada palabra incluida en los conjuntos

antes mencionadas se almacenan todas sus formas flexionadas y los sinónimos con sus diferentes flexiones. La ventaja de esta representación radica en que el tratamiento de la ambigüedad semántica, expresada en la sinonimia, se reduce a partir de que los sinónimos considerados son aquellos válidos en el contexto de la pregunta.

Una característica importante en este sistema para su consideración es que la representación y mantenimiento de las preguntas almacenadas no se realiza de forma automática. Aunque esto permite un mejor resultado en la tarea final, en la práctica, limita su aplicación como sistema independiente y escalable.

1.3 Herramientas Computacionales Necesarias para el Desarrollo de un Sistema de Consulta en Lenguaje Natural a ficheros de Preguntas Frecuentes

En las dos propuestas analizadas anteriormente "FAQ Finder" y "Automated FAQ Answering" se aprecian dos acercamientos cuyas diferencias afloran si intentamos responder la siguientes preguntas: ¿Qué información se representa? ¿Qué criterio se toma para el matching entre las preguntas?

Las limitantes de la segunda propuestas están en lo que señalamos anteriormente de que, la representación y mantenimiento de las preguntas se realiza de forma manual por un especialista. Pero a su favor tiene que obtiene los mejores resultados. El primer modelo realiza el análisis de la pregunta de entrada automáticamente, pero el referido análisis solo se limita a eliminar las palabras contenidas en la lista de palabras irrelevantes estadísticamente (*stop-words*) y asociarle un grado de relevancia a cada término a partir de la frecuencia de aparición en el fichero de pares Pregunta-Respuesta. El procesamiento lingüístico, utilizado para el matching entre las preguntas, se reduce a la búsqueda de la distancia entre dos términos encontrando el punto más cercano de coincidencia en sus

árboles de hiperónimos. En ningún momento se realiza un análisis de la función que realiza la palabra dentro de la pregunta. Las razones expuestas son la necesidad de lograr un nivel aceptable entre la cobertura y el tiempo de respuesta. Es importante destacar que en la bibliografía revisada no queda claro el proceso que se lleva a cavo en la primera propuesta. Lo que exponemos es el resultado del análisis directo del trabajo [Burke et. al. 1997], pero la duda surge a partir de la referencia hecha a este trabajo en la publicación [Sneiders] donde se describe como un sistema que además identifica en la oración los sintagmas nominales y verbales.

Nuestro objetivo va a ser determinar, en primer término, el grado de relevancia del término dentro de la pregunta a partir de la función que realiza y lograr un indexado de la pregunta de modo que se codifique esa información. En general proponemos lograr un nivel de representación similar al propuesto en el segundo enfoque pero de forma automática. Los retos más importantes en esta tarea estarían en lograr restringir el conjunto de sinónimos de una palabra al contexto de la pregunta en que se expresa y lograr indexar las preguntas de modo que se codifique la función que realiza cada palabra dentro de la pregunta.

Para lograr un nivel de análisis de las preguntas como el deseado se necesita contar con varias herramientas computacionales para el procesamiento de Lenguaje Natural:

- Analizador Morfológico
- Desambiguador Morfosintáctico
- Analizador Sintáctico Superficial
- Ontología
- Indexador

Estas herramientas son descritas a continuación.

1.3.1 Analizador Morfológico

La responsabilidad de un analizador morfológico es el reconocimiento de cada una de las palabras, así como sus características desde el punto de vista morfosintáctico. Nuestro objetivo en particular: obtener de cada palabra su primitiva (lematización), las posibles categorías gramaticales que admite (etiquetación) y reconocer las flexiones presentes en cada una de sus interpretaciones.

En este tema no nos detendremos mucho porque en el capítulo número tres se aborda en profundidad y en el capítulo cuatro se expone una propuesta de implementación.

1.3.2 Desambiguador Morfosintáctico

La responsabilidad de un analizador morfológico se limita a enunciar las posibles variantes de análisis que asimila una palabra. Este resultado se obtiene sin tener en cuenta el contexto en que se encuentra expresada. Por ejemplo: para la palabra "camino" se podrían obtener dos interpretaciones. Una como sustantivo masculino singular y otra como la conjugación en primera persona del singular del presente del modo indicativo del verbo caminar. Ahora bien. ¿Cómo saber cuál escoger en cada momento? La respuesta a esta pregunta se encuentra en las palabras que la preceden o suceden, palabras, a las que el analizador morfológico no tiene acceso.

La responsabilidad de discriminar entre las posibles variantes de interpretación morfosintáctica de un término en un contexto específico, una oración, corresponde a los desambiguadores morfosintácticos o etiquetadores "Part of Speech Tagger" (en inglés).

Tradicionalmente el problema de la desambiguación morfosintáctica se ha enfocado desde dos puntos de vistas:

- Lingüístico
- Estadístico

El enfoque lingüístico busca representar el problema utilizando modelos lingüísticos que representen toda la información necesaria para desambiguar. Habitualmente estos se implementan a partir de un conjunto de reglas. Estos sistemas tienen la ventaja de que permiten especificar de forma explícita información lingüística y pueden representar gran cantidad de conocimiento y fenómenos complejos. Pero, no obstante a sus innegables ventajas, su desarrollo es realmente costoso al necesitarse de una gran cantidad de información obtenida a partir de la observación y el estudio profundo de los fenómenos generales y casos particulares en su uso, además de que su utilidad se limita al lenguaje para el cual se desarrolle.

Como variante al enfoque lingüístico aparecen los sistemas que abordan el problema a partir de la información estadística que aportan las frecuencias de ocurrencia de los n-gramas en los corpus ya sean etiquetados (entrenamiento supervisado) o sin etiquetar (entrenamiento no supervisado) obteniendo gran robustez y cobertura. Múltiples son las ventajas de este enfoque que van desde la simplicidad del proceso de desarrollo hasta la portabilidad de un lenguaje a otro. Como ejemplos representativos de este enfoque están los sistemas que emplean el formalismo conocido como Modelos Ocultos de Markov, HMM (por sus siglas en inglés).

1.3.3 Analizador Sintáctico Superficial (Shallow Parsing)

Cuando nos enfrentamos a la tarea de realizar el análisis sintáctico de oraciones en lenguaje natural resulta vital determinar el grado de profundidad que requerimos en el análisis. En nuestro caso, cómo en el de la mayoría de las aplicaciones, sería deseable obtener como resultado del análisis una descripción robusta de la sintaxis de la pregunta analizada, más sin embargo, la realidad de los sistemas que realizan este tipo de análisis es que tienen muy poca precisión fundamentalmente debido a que para tal nivel de profundidad se requiere tanto de información sintáctica como semántica lo que, adicionalmente, limita su velocidad. No obstante en dominios restringidos donde el

contexto tanto semántico como léxico puede ser más local, estos análisis, logran prestaciones aceptables.

Considerando que la aplicación no impone restricciones de dominio y que, además, lo que nos interesa de los términos presentes en la oración es determinar su grado de relevancia y la función que realiza dentro de ella, sería suficiente con un análisis parcial tal como se considera en el sistema FAQ Finder.

La idea principal es reconocer piezas o fragmentos (tradicionalmente llamados *chunks*) a partir de la información puramente morfosintáctica de las palabras aportada por el analizador morfológico en un paso previo. Un ejemplo de una definición de chunk es la que se propone en [Torruella et. al] para un chunk nominal: sn = [(det.) + (sadj.) + núcleo + ({sn | sadj. | sp[de +sn]})] donde (det.) es un determinante, (sadj.) sintagma adjetivo y (sp) sintagma preposicional.

1.3.4 Módulo para el tratamiento de Sinónimos e hiperónimos (Ontología)

Pondríamos pensar el módulo de tratamiento de sinónimos e hiperónimos como una ontología en la que las relaciones se restringieran a la sinonimia e hiperonimia. Restringir el conjunto de relaciones a estas dos planteadas es el resultado de las conclusiones encontradas en los trabajos que tomamos como base para esta investigación. Pero, no obstante, sería interesante indagar en las ventajas de incluir otra relación de gran uso como la meronimia (es parte de).

Las redes semánticas han sido una forma de representación del conocimiento ampliamente utilizada, pero el hecho de limitar el conjunto de relaciones hace posible pensar en una estructura más simple como la propuesta en [Burke et. al 1997] para la hiperonimia.

Aunque sin aportar una solución, en este apartado queremos hacer notar una consideración importante fruto del estudio realizado: para lograr un nivel de representación como el propuesto en [Sneiders] es importante intentar no considerar la

distancia entre dos palabras de forma aislada, sino restringida al contexto en que se expresan, o sea, lograr una medida de la cercanía entre dos términos en donde influyan las palabras que se encuentran en la oración. Más explícitamente, considerar como sinónimos he hiperónimos válidos de una palabra sólo aquellos que sean compatibles con el contexto en que ella se expresa.

1.3.5 Indexador

El volumen de información que potencialmente puede llegar a manipular el sistema hace necesaria una representación adecuada que permita el acceso y la extracción eficiente de las características que definen las preguntas almacenadas. Ahora bien. ¿Qué características? ¿Cómo representarla? ¿Cómo acceder de forma eficiente?

Aunque Emilio Rodríguez Vázquez de A., en su trabajo [Rodríguez E.] concluye planteando que: "las aportaciones de las Técnicas de Procesamiento de Lenguaje Natural a la Recuperación de Información no han sido las que se esperaban" y que "específicamente la indexación sintáctica, en general, no ha producido mejoras". En nuestro caso se hace imprescindible la representación de información lingüística asociada a las palabras. ¿Razones? Básicamente radican en que los pares pregunta-respuesta, debido al reducido número de palabras que contienen, no aportan información suficiente a los métodos estadísticos tradicionales.

En estos momentos de la investigación no nos hemos planteado una estrategia definitiva aunque a primera impresión parece suficiente con codificar además de las palabras el sintagma al que pertenece, ya sea nominal o verbal y además diferenciar el núcleo de los complementos.

Conclusiones

En este capítulo hemos introducido el problema que nos ocupa y, de modo superficial, descrito las herramientas básicas necesarias para su desarrollo. Tomamos como referencia los trabajos: FAQ Finder y Automated FAQ Answering que plantean dos aproximaciones diferentes al problema. Nuestra propuesta de solución se acerca un poco más a la primera, tratando de profundizar en las relaciones sintácticas que se establecen entre los términos de la oración con el fin de determinar el grado de relevancia, logrando así una representación que permita un análisis con cierto grado de parecido al Prioritized Keyword Matching expuesto en Automated FAQ Answering. Como peculiaridad proponemos además realizar un indexado de las preguntas almacenadas en la BC para acelerar el proceso de Matching con la pregunta de entrada.

Capítulo 2.

Consideraciones sobre la Morfología de la Lengua Española

Introducción

El español es una lengua con una flexión bastante rica, por ejemplo en la flexión verbal, para tiempos simples, hay alrededor de 61 formas flexivas, 6 formas para el duplicado subjuntivo pasado imperfecto. Adicionalmente hay 45 formas para tiempos compuestos, resultando en 112 formas posibles para cada verbo [Velásquez et. al], lo que hace del estudio de su morfología un problema no trivial.

En este capítulo realizaremos un estudio de los principales aspectos lingüísticos que caracterizan nuestra lengua: procesos de formación de palabras, constitución fónica de los morfemas gramaticales con las principales adaptaciones fonéticas que se realizan durante su concatenación. Realizamos una descripción de los modelos utilizados tradicionalmente por los lingüistas para la descripción de la morfología y exponemos algunas peculiaridades de la flexión nominal y verbal. Finalmente describimos las clases de palabras del español atendiendo a la función que realizan dentro de la oración.

2.1 Lexicogenesia Nominal.

Todas las lenguas romances han aumentado el caudal de sustantivos y adjetivos modificando de diversas formas las palabras latinas para crear otras nuevas. Los procedimientos seguidos son diversos y entre ellos se encuentran los siguientes:

- Habilitación

  Mediante el cambio de la función originaria de la palabra. Por ejemplo *la sustantivación* al anteponer un artículo: el comer, el valiente, etc. Encontramos

también el paso de sustantivos propios a comunes: lázaro, quijote, poder, placer, etc. Aunque sin la vitalidad anterior, la adjetivación de sustantivos y participios tiene una gran variedad de ejemplos: chico (pequeño), hondo (profundo). Por otra parte, un gran número de animales son usados como adjetivos para mostrar metafóricamente una cualidad que resalta de una persona: lechuzo, lince, burro, zorro, tortuga, etc.

- Derivación Sufijal

  Es el recurso más abundante en la formación de palabras nuevas. Para que en romance un sufijo pueda vivir productivamente, esto es, producir nuevas palabras necesita ser tónico. La casi totalidad de los sufijos en romance son procedentes del latín. Algunos revisten doble forma por haberse introducido por el doble camino de la tradición oral y la escrita.

- Derivación Prefijal

  Al contrario que los sufijos, los prefijos en romance son átonos; la mayoría son de origen latino; pueden ser preposicionales o adverbiales, según la función secundaria que aporten a la nueva palabra que, como nueva unidad léxica, tiene carácter nominal: anteiglesia, entreacto, sotavento, deshora, rebueno, bizcocho, etc.

2.2 Constitución Fónica de los Morfemas Gramaticales

El español actual está constituido de tal suerte que los morfemas gramaticales tienen una constitución fónica relativamente determinada según la categoría a la cual pertenezca. El morfema de número terminal, es consonántico; el de género, que le precede, es vocálico. El sufijo o el infijo es del tipo "vocal + consonante", terminando frecuentemente el lexema en una consonante. En cuanto al prefijo, termina en vocal o en consonante débil (implosiva) a fin de permitir que la consonante inicial del lexema permanezca inicial de la sílaba.

La mayoría de las palabras castellanas corresponden a una fórmula que presente estas características:

Cv(c)		CvC(vC)		vC		v		C
(prefijo)		(lexema)		(inf., suf.)		(G)		(Núm.)
		mes		*it*		*a*		*s*
		mord		*isc*		*o*		*s*
des		*templ*		*anz*		*a*		

2.2.1 Adaptaciones Fonéticas en la Concatenación de Morfemas

Cuando las junturas entre los morfemas son poco viables, tienen lugar adaptaciones fonéticas (asimilaciones, intervenciones...), algunas de las cuales enumeramos a continuación:

- *Supresión de la vocal final átona:* al concatenar los sufijos se elimina la vocal final del término base en el caso de sustantivos y adjetivos. Ej: *arena + oso > arenoso.*
- *Eliminación de cacofonías:* en ocasiones, al concatenar el sufijo a la raíz obtenida mediante el proceso anterior, dos vocales iguales quedan adyacentes. Ambas se fusionan para eliminar la cacofonía resultante.
- *Vocal temática:* en el caso de que el término primitivo sea verbo, basta con comprobar si acaba en *–ar/ -er/ -ir/ -ír* para conocer la vocal temática y así ser tenida en cuenta, por ejemplo, a la hora de escoger la variante alomórfica a utilizar. Una muestra es el caso de *–miento/ -amiento/ -imiento/ -mento,* donde *–amiento* sólo se emplea con vocal temática "a" e *–imiento* con las vocales "e" e "i".

- *Monoptongación de la raíz diptongada:* se sustituye el diptongo por la forma pertinente. Se considera la monoptongación de *ie* en *diente* > *dental* y de *ue* en *fuerza* > *forzudo.*

- *Cambio en la posición del acento:* puesto que los sufijos producen generalmente un cambio en la acentuación, dicha situación debe ser considerada, ya que puede conllevar cambios ortográficos debidos a la aparición o desaparición de tildes. La práctica totalidad de los sufijos son tónicos, con lo que es inmediato saber si debemos introducir o eliminar una tilde aplicando las reglas ortográficas pertinentes.

- *Reglas ad-hoc:* son ajustes varios tales como modificaciones en la consonante final de la raíz en cambios como la derivación de *concesión* a partir de *conceder.* Estos casos se resuelven mediante reglas *ad-hoc*, es decir, que operan para un sufijo dado. Frecuentemente vienen dados por la presencia de fonemas dentales /δ/ o /t/.

- *Mantenimiento de los morfemas consonánticos finales:* conociendo el fonema podemos deducir la ortografía final. Por ejemplo, la *z* en *cerveza* corresponde a /θ/ y por consiguiente la *c* de cervecería corresponde también a /θ/ y no a /k/. Los fonemas y cambios cubiertos son:

  /k/ c → qu

  /γ/ g → gü

  /γ/ g → gu

  /θ/ z → c

  /θ/ c → z

2.3 Modelos Lingüísticos

2.3.1 Elemento y Distribución

En este modelo el acento está en tres aspectos que comparten todas las lenguas.

a) hay elementos que figuran en diferentes contextos y que, gracias a esto, pueden aislarse de la cadena sonora.
b) los enunciados transcurren en el tiempo, es decir, hay sucesión de elementos, lo cual, en la representación gráfica se expresa en forma lineal y se ha llamado comúnmente la linealidad del lenguaje.
c) todas las lenguas muestran secuencias típicas de clases de morfemas.

De acuerdo con este modelo, en vez de partir de un léxico de palabras, se tendría un inventario de todas las construcciones posibles como palabras, completadas con las listas de los morfemas que pueden aparecer en cada posición. El significado de las palabras quedaría explicitado por la combinación de los significados de los morfemas y los valores de las construcciones o relaciones características al interior de la construcción. Una palabra como "*anticonstitucional*" se supondría formada así: *anti + constitu + ción + al.*

En contra de este modelo tenemos los siguientes elementos:

a) El significado de las palabras varía a lo largo de la historia, lo que muestra que tienen vida como unidades autónomas.
b) Los afijos evolucionan semánticamente, proceso que sólo puede darse al interior de las palabras, concomitantemente con la evolución de estas.
c) La aparición de formas nuevas de afijos a través del fenómeno conocido como resegmentación.

d) En un análisis en morfemas ya aislados de su contexto, éstos se tendrían que representar como elementos polisémicos y polifuncionales pues no hay correspondencia entre forma, significado y función.
e) La falta de autonomía y fijeza semántica del afijo.

Un hablante que para expresar encadenara morfemas, tendría que añadir, además, al inventario de éstos, tal como lo hace el análisis *Elemento Distribución*. Un complejo juego de reglas morfofonológicas y fonológicas para producir palabras aceptables. Lo menos que se puede decir es que hablar por morfemas en vez de por palabras sería muy costoso.

Para el Español el modelo más económico de expresar restricciones sobre la coocurrencia de sufijos es haciendo referencia a la clase de palabras que determina un uso.

No sería económico enumerar los grupos de sufijos una y otra vez bajo las posiciones donde pueden aparecer. Es preferible expresar las condiciones de uso haciendo referencia directamente a las clases gramaticales, aunque sea necesario especificar luego subclases de afijos y condicionamientos entre morfemas concretos:

V → St exploración y fundamentación

V → Aj adherible y volatilizable

St → V volantear y fundamentar

St → Aj campal e institucional

Aj → V inmunizar, inutilizar, institucionalizar

Aj → Av comúnmente y sensiblemente

Av → V adelantarse

Av → Aj abajeño, fuereño

Donde: V = verbo, St = sustantivo, Aj = adjetivo, Av = adverbio.

2.3.2 Elemento y Proceso

Este modelo, al mismo tiempo que se basa en uno de los aspectos claves de *Elemento Distribución* – en el hecho de que todas las lenguas tienen elementos recurrentes – enfoca más la relación entre construcciones parcialmente semejantes que la cohesión distribucional o frecuencia de coocurrencia de los constituyentes últimos.

En *Elemento y Proceso* se postula una relación de procedencia o derivación de la forma compleja con respecto a la inmediata más simple; es decir, se considera que una forma compleja consiste en una forma contenida en ella que se ha sometido a un proceso. Aquí *constitucional* se explicaría como derivada mediante la anteposición de un afijo (anti-) a la palabra *constitucional* y esta última como derivada de *constitución* mediante sufijación (-al), etc. y el significado por el tipo de relación entre constituyente inmediato y el sufijo.

Corrobora la realidad psicológica de esta relación la capacidad de los hablantes para obtener de las palabras derivadas otras menos complejas, es decir, su capacidad para las formaciones retrógradas o regresivas. Esta facultad, documentada en formaciones regresivas no etimológicas, obliga a representar la capacidad derivacional en una gramática, no tanto como relación de procedencia, sino como relación entre términos del léxico.

Los principales aspectos que limitan a este modelo son los siguientes:

a) Resulta difícil, muchas veces, decidir acerca de qué forma o formas contenidas reconocer debido, entre otros, a que muchos supuestos procesos producen cambios formales en elementos integrados en la forma mayor (como la pérdida de desinencias que permitirían identificar la clase primitiva).
b) Se infieren primitivas que no "existen" (que no están y no han estado en uso) en la creencia de estar ante palabras ya consagradas. El DRAE está lleno de tales creaciones progresivas y regresivas, aunque muchas veces justifica lo que parecen ser creaciones suyas remitiéndose al latín, como cuando da entradas a los

términos *amigar* y *enemigar*. Éstas se pueden obtener en sincronía del siguiente modo: *amigable: amigar* con base en la experiencia de que los adjetivos en *–ble* suelen acompañar a verbos, y, con la conciencia de la oposición *amigo/ enemigo*, el antónimo *enemiga*.

Estas restricciones hacen proponer un juego de reglas, en vez de una sola, para los procesos relacionados con determinados afijos.

Así para *–izar*, por ejemplo se daría:

$((X)_{St}$ izar$)_V$ estilizar, valorizar y $((X)_{Aj}$ izar$)_V$ catolizar, relativizar

y para *–ción:*

$((X)_V$ ción$)_{St}$ institución

$((X)_{Aj}$ ción$)_{St}$ inanición

$((X)_{St}$ ción$)_{St}$ sudoración

2.3.3 Palabra Paradigma

Este modelo usado tradicionalmente para representar flexión, despliega todas las formas flexionadas de una palabra en cuadros ejemplares o paradigmas. Al conocerse una forma (por ejemplo *alunizaron*) y el paradigma que es representativo para esa forma (1ra conjugación) se pueden construir todas las demás formas de la palabra en analogía con las que se encuentran en ese paradigma.

De manera similar podrían representarse las posibilidades derivacionales de una clase de palabras ejemplificándolas en las actualizaciones diversas de un lexema tipo.

El conocimiento que representa el modelo Palabra y Paradigma puede parecer en principio el mismo del que expresa Elemento y Proceso: *vendible* se relaciona con VENDER, o sea, $((X)_V$ ible$)_{Aj}$, pero con un mayor grado de sistematización, ya que agrupa todos los postverbales en un cuadro, todos los postnominales en otro, et.

Sin embargo, si se piensa en cuadros de elementos concretos, agrupados en torno a su primitiva o, mejor aún, agrupados, en familias de palabras, se observan varias diferencias importantes de este modelo con respecto al anterior y que aportan algo a la compresión del fenómeno.

a) La idea de que una palabra derivada se construye sobre el modelo de otra derivada y no a partir de una regla que selecciona las primitivas.

b) La introducción del concepto de analogía.

c) La información que aporta un cuadro que permite percibir oposiciones múltiples; es decir, que deja ver la posibilidad de relacionar o derivar cada forma de todas las demás y no impone una vía única.

Con este modelo queda explicado, o por lo menos previsto el uso de un afijo con bases diferentes y la posibilidad de que un derivado no se forme exactamente con base en una primitiva, sino bajo la influencia de *varios* miembros de la familia de palabras. Así por ejemplo, *pedregoso* puede haberse formado a partir de *piedra*, pero bajo la influencia de *pedregal*.

En otras palabras, este modelo muestra una flexibilidad comparable a la de los hablantes.

A partir de este planteamiento se puede proponer que en los casos en que la mayoría de las palabras derivadas con un determinado afijo pertenecen a familias de estructuras similares, las abstracciones que hagan los hablantes también sean similares, con el resultado de que se tiende al establecimiento de una regla del tipo Elemento y Proceso; y que aquellos derivados que muestran el mismo afijo, pero pertenecen a familias de estructura muy variada den lugar a abstracciones variadas o que vean bloqueadas las posibilidades de abstraer pautas con la consiguiente improductividad del afijo.

Esta podría ser la diferencia entre palabras en *–ble* o en *–idad* y las formadas en *–orio* (perentorio, ilusorio, notorio) para poner un ejemplo.

Como limitación para esta representación está el hecho de que no se encuentra una palabra para la cual se hayan realizado todas las posibilidades derivacionales, sino que lo típico es que se realicen diferentes derivados para diferentes palabras de la misma clase.

2.4 Particularidades de la Flexión Nominal y Verbal.

El español como lengua flexiva, posee un gran número de procesos morfológicos, particularmente los no concatenativos. A continuación relacionamos algunos de ellos, a partir de los datos publicados en [Velásquez et. al].

Un paradigma verbal muy complejo. Para tiempos simples, hay alrededor de 61 formas flexivas, incluyendo el duplicado subjuntivo pasado imperfecto (6 formas). Si agregamos las 45 posibles formas para tiempos compuestos, hay 112 formas flexivas posibles para cada verbo.

La frecuente irregularidad de raíces y terminaciones verbales. Verbos muy comunes, como *tener, poner, poder, hacer, etc.* Tienen hasta 7 raíces: *hac-er, hag-o, hic-e, ha-ré, hiz-o, haz, hech-o.* Aunque el 85 % de los verbos en español son regulares.

Huecos en algunos paradigmas verbales. En los llamados verbos defectivos algunas formas se pierden o simplemente no se usan. Por ejemplo, los verbos meteorológicos como *llover, nevar, etc.*, son conjugados sólo en tercera persona del singular. Otros son más peculiares como *abolir* que falla en primera, segunda y tercera persona del singular y tercera del plural del presente indicativo, en presente del subjuntivo y en la segunda persona del singular de la forma imperativa. En otros verbos, los tiempos compuestos se excluyen del paradigma, como *soler.*

Participios pasados duplicados. Una cantidad de verbos tienen dos formas alternas, ambas correctas, como *impreso, *imprimido (no en cuba).*

Algunos sustantivos y adjetivos presentan formas alternativas correctas para el plural: *bambú → bambús, bambúes.*

Hay un pequeño grupo, (3%), de sustantivos invariantes con la misma forma para el singular y el plural (ej. *crisis*).

El 30 % de los adjetivos presentan la misma forma para el masculino y el femenino (ej. *azul*).

Existen los *singularia tantum,* que sólo usan el singular como *estrés,* y los *pluralia tantum* que sólo se usan en plural como en *matemáticas.*

2.4 Clases de Palabras

Antes de introducirnos en la descripción de las clases de palabras de nuestra lengua veamos algunos conceptos y elementos introductorios:

Los Monemas: El monema es la unidad de la primera articulación del signo lingüístico. Esto quiere decir que cualquier palabra puede ser dividida en unidades más pequeñas, dotadas de un significante y un significado. Así pues, una palabra podrá estar constituida por uno o más monemas.

Los Lexemas: No todos los monemas significan de la misma forma. Distinguimos, en primer lugar, los lexemas, que son los que aportan el significado fundamental de la palabra; por ello se dice que son como la **raíz** de la palabra.

Los Morfemas: los morfemas son monemas que desempeñan dos funciones: añaden nuevos matices a la significación básica del lexema y sirven para relacionar unos lexemas con otros.

Clases de morfemas.

- **Morfemas dependientes o trabados**: son aquellos que necesitan unirse a un lexema para tener significado.
- **Morfemas dependientes derivativos o afijos.**

  Modifican el significado básico del lexema.

  -sufijos: si se colocan después del lexema.

  -prefijos: si preceden al lexema.

- interfijo: Se sitúan entre los prefijos y sufijos para evitar la cacofonía (sonido desagradable) entre dos sonidos. No tienen ninguna significación.

Morfemas dependientes gramaticales. Frente a los afijos, otros morfemas no sólo modifican el significado básico del lexema, sino que además nos sirven para relacionar esa palabra con otras. Ocupan siempre la posición final de la palabra y sirven para expresar los accidentes gramaticales. Se llaman flexivos porque nos muestran las diversas posibilidades o flexiones de una palabra. Los morfemas flexivos del verbo reciben el nombre especial de desinencias. La información que ofrecen es de tipo **gramatical**, como el género, el número, la persona, el modo, etc.

Morfemas independientes o libres: son aquellos que no necesitan ir unidos a ningún lexema, sino que forman por sí solos una palabra. Son morfemas independientes los determinantes, las preposiciones y las conjunciones.

2.4.1 Clases de Palabras según sus Monemas

Atendiendo a los distintos tipos de monemas que pueden componer una palabra, estas pueden clasificarse en distintas clases:

- **simples**: un sólo lexema o un morfema independiente. Ej. *cocodrilo*: *lexema*; **en: morfema independiente.**
- **derivadas**: un lexema más morfemas derivativos o afijos. Ej. ***avion**-eta:* ***lexema**+morfema derivativo*
- **compuestas**: dos o más lexemas. Ej. ***moto**carro*: **lexema**+*lexema*.

Al formar una palabra compuesta, tenemos que tener en cuenta que unimos las palabras y, en parte, los significados, pero el significado final no equivale a la suma de los significados parciales, sino que se refiere a una realidad nueva. Se suman significados lingüísticos, no referentes.

También hay que tener en cuenta que muchas palabras compuestas se han formado partiendo de **lexemas latinos y griegos**, para referirse a nuevas realidades que no existían ni en tiempos de los griegos ni de los romanos. Ej. ***tele**visión*: **lexema**+*lexema*.

Parasintéticas: Hay dos conceptos diferentes de parasíntesis, que no guardan relación entre sí.

- Dos o más lexemas más morfemas derivativos o afijos. Es decir, es la suma de composición más derivación. Ej. balon*cesti*sta: lexema+*lexema*+sufijo.
- -Prefijo más sufijo que se necesitan obligatoriamente. No existe la palabra formada por el prefijo más el lexema, o el lexema más el sufijo. Ej. *en*granaje: *prefijo*+lexema+*sufijo*, sin que exista *engrano ni *granaje.

2.4.2 La Base Significativa y los Morfemas Constituyentes

Podemos distinguir las distintas clases de palabras atendiendo al modo de significar de las palabras y a sus morfemas característicos. Por ello conviene distinguir entre la base significativa y los morfemas gramaticales.

- **Base significativa:** el significado resultante de la suma del significado del lexema más las modificaciones expresadas por los morfemas derivativos o afijos, pues éstos añaden matices al significado.
- **Morfemas gramaticales:** Podemos distinguir las diferentes clases de palabras por los morfemas que las caracterizan. Veremos que cada clase de palabra tiene unos morfemas gramaticales característicos, que siempre están presentes: son los morfemas constituyentes de las diferentes clases de palabras.

En español, distinguimos nueve clases de palabras: sustantivos, adjetivos, determinantes, verbos, adverbios, pronombres, conjunciones e interjecciones.

2.4.3 Los Sustantivos

El sustantivo o nombre está compuesto por una base significativa más el morfema constituyente de número. El género no es propiamente un morfema, ya que en muy pocos

casos existe variación genérica, como veremos más adelante. Podríamos decir que el género de los sustantivos se incluye dentro del significado de los nombres.

2.4.3.1 Base significativa.

Los sustantivos o nombres son las palabras que empleamos para designar los seres, cosas e ideas, ya sean reales o fingidas. Además la base significativa nos proporciona el género de los nombres. Ej. ***casa***: "lugar donde se vive" + "género femenino"

2.4.3.1.1 El género

El género indica mayoritariamente el **género gramatical** de los nombres. En los sustantivos sólo hay dos géneros: masculino y femenino. Ello quiere decir que todos los nombres tienen uno u otro, pero se trata de género gramatical, no de género "real". Sólo en los casos en que el nombre se refiere a personas o animales (y no siempre) el género gramatical se corresponde con el género real: *niño / niña - burro / burra.* En los nombres que designan a los animales es frecuente la indistinción de género "real": la misma palabra designa tanto al macho como a la hembra. Son los **nombres epicenos**. Gramaticalmente, sólo poseen un género. Si se quiere especificar el sexo, debemos añadir las palabras "macho" - "hembra": *La perdiz macho / La perdiz hembra.* También es muy frecuente designar con palabras distintas el sexo de los seres animados. En ese caso, decimos que el género se expresa con **heterónimos**: *Caballo / yegua - Hombre / mujer - Toro / vaca.*

Finalmente, hay palabras que se pueden usar indistintamente en masculino o femenino, pues la lengua permite tal vacilación genérica, sin que cambie el significado: *el / la mar - el / la centinela - el / la color.*

En ocasiones, el cambio de género de un mismo nombre implica **cambio de significado**: *el / la cometa, el / la corte, etc.*

2.4.3.2 Morfemas constituyentes: el número

El número indica la oposición entre uno y varios, es decir, singular y plural. Además tiene unos morfemas siempre fijos:

	Singular	Ejemplo	Plural	Ejemplo
Morfemas	-Ø	el lunes	-Ø	los lunes
	-Ø	ese alumno	-*s*	esos alumno*s*
	-Ø	un candil	-es	cuatro candil*es*

Tabla 1: Mofemas constituyentes de número en los sustantivos

No siempre la oposición singular / plural significa uno/ varios. El singular, a veces, equivale a toda una especie. Ej. *El deber del alumno es estudiar,* donde ***alumnos*** equivale a "todos los alumnos".

Algunas palabras se utilizan siempre en singular. Se denominan **singularia tantum**. Ej. *la tez, el nadir, la sed.* De igual forma, ciertas palabras se emplean sólo en plural, o **pluralia tantum**. Ej. *víveres, nupcias, andas.*

En ocasiones podemos utilizar indistintamente el singular y el plural para referirnos al mismo objeto: **Tráeme esos *pantalones* / Tráeme ese *pantalón*.**

Los nombres **abstractos**, cuando se usan en plural, pasan a ser concretos. Ej. *La **maldad** de los profesores es increíble/ Las **maldades** de los alumnos son increíbles.*

E incluso pueden dejar de serlo en singular, sobre todo en los usos poéticos. Ej. *La **avaricia** rompe el saco.*

2.4.3.3 Otros morfemas facultativos: los apreciativos

Hay una clase de **afijos** (puede llevar, además, muchos otros) que se ligan al nombre: son los **apreciativos**, es decir los afijos que expresan la actitud del hablante ante el nombre. Hacemos referencia al tamaño de los objetos, aunque esto se cumple en muy pocas ocasiones. Generalmente aluden a un hecho subjetivo, es decir, expresamos admiración, estima, cariño o aversión hacia el objeto designado por el nombre.

Se dividen en:

- **Aumentativos**: con ellos aumentamos nuestra consideración del tamaño, real o ficticio, del objeto señalado por el nombre. Y también expresamos nuestra admiración y sorpresa ante el objeto. Ej. ***Cochazo*** puede ser tanto un "coche muy grande" como un "coche muy apreciado por nosotros, aunque su tamaño sea pequeño".
- **Diminutivos**: disminuyen la magnitud del lexema. En la mayor parte de los casos, el diminutivo tiene un valor afectivo: expresamos con ellos cariño y estimación. Ej. ***Papaíto*** no significa `papá pequeño', sino que mostramos nuestro cariño hacia nuestro padre.
- **Despectivos**: modifican el lexema con una información de desafecto. Ej. **Cas*ucha*, poeta*stro*.**

En ocasiones, los apreciativos han dejado de cumplir la función que tenían en un principio. Se han convertido en afijos derivativos normales. Pierden su valor apreciativo. Ej. En ***sillita*** el sufijo tiene valor diminutivo, pero no ocurre lo mismo en ***sillón***, donde el sufijo se convierte en morfema derivativo, pues no significa `silla grande'.

2.4.4 Los Adjetivos

El adjetivo calificativo está compuesto por una base significativa más tres morfemas constituyentes: **grado**, **género** y **número**. Además se diferencia del nombre por su posibilidad de convertirse en otra clase de palabra, el adverbio, al añadirle el morfema trabado ***-mente***.

2.4.4.1 Base significativa.

El adjetivo es una palabra que no tiene independencia lingüística. Necesita la existencia de un nombre para existir, pues expresa siempre una cualidad del sustantivo al que acompaña. Es un elemento "adjuntado" al sustantivo, del que expresa cualidades, y siempre irá referido a éste, sea cual sea su función sintáctica. Su significado dependerá de

su lexema, que siempre expresará cualidades. Ej. *Las paredes blancas de la clase. Estos alumnos son trabajadores.*

2.4.4.2 Morfemas constituyentes.

El adjetivo tiene tres morfemas constituyentes: grado, género y número.

2.4.4.2.1 El grado

El adjetivo puede ser "graduado"(excepto los **adjetivos calificativos determinativos)**, puesto que expresa cualidades y éstas admiten gradación. Existen tres grados en el adjetivo: positivo, comparativo y superlativo.

- **Grado positivo.**

  Es el adjetivo tal y como se nos presenta a nosotros para ser utilizado, sin que exprese ningún tipo de graduación. El morfema, por tanto, es siempre Ø. Ej. *Los árboles altos del jardín lejano están deshojados*

- **Grado comparativo**

  Se establece una comparación de cualidades entre dos o más sustantivos, o se comparan dos cualidades de un mismo sustantivo:

  Estos árboles están más deshojados que aquellos

  Juan es más trabajador que inteligente

- **Grado superlativo.**

  Es el grado del adjetivo en su mayor potencialidad. Por ello también se le conoce como superlativo absoluto. Se pondera al máximo la cualidad del adjetivo. Hay varios procedimientos para formar el superlativo:

  - Mediante el morfema gramatical dependiente **-ísimo/a/os/as** u otros parecidos, pero de origen culto, como **-érrimo/a/os/as.** Ej. *Fuimos a un lugar lejan**ísimo.** Ese señor que pide limosna es paup**érrimo.***

  - Mediante otros morfemas gramaticales, generalmente antepuestos al lexema del adjetivo, como **re-, requete-, super-.** Ej. *Vimos un coche **super**rápido. Pasaban unas chicas **requete**guapas.*

- Mediante el morfema libre **muy**, antepuesto al adjetivo: Vimos un coche **muy** rápido. Pasaban unas chicas **muy** guapas.

Nota: Además hay formas que expresan, sin morfemas especiales, los diferentes grados de un adjetivo. Proceden del latín. Son muy pocos casos los que nos quedan y, en su mayoría, han perdido su valor comparativo o superlativo.

Positivo	Comparativo	Superlativo
grande	mayor	máximo
pequeño	menor	mínimo
bueno	mejor	óptimo
malo	peor	pésimo

Tabla 2: Otras formas de expresar el grado en los adjetivos

2.4.4.2.2 El género.

El género del adjetivo depende del sustantivo al que se refiera. Se dice que el adjetivo concuerda con el género del sustantivo. Se presenta como un morfema, pues en todos los adjetivos es posible la variación de género, aunque en muchos casos el morfema sea Ø. Por ello se habla de adjetivos de **una terminación** (la variación de género no implica variación en la forma) y de **dos terminaciones**. Por tanto, para reconocer el género del adjetivo siempre tendremos que tener en cuenta al sustantivo al que va referido.

		Morfema		Morfema
Femenino	La casa **verde**	-Ø	La casa **alt*a***	-a
Masculino	El pino **verde**	-Ø	El pino **alt*o***	-o
	Una terminación		Dos terminaciones	

Tabla 3: Los morfemas de género en los adjetivos

2.4.4.2.3 El número.

Al igual que el género, el número del adjetivo viene impuesto por el sustantivo al que hace referencia. Así podemos decir que el adjetivo concuerda en género y número con el sustantivo. Sus morfemas son los mismos que los del sustantivo.

		Morfema
Singular	La casa verde/azul/beis	-Ø
Plural	Las casas **beis**	-Ø
Plural	Las casas **verde*s***	-s
Plural	Las casas **azul*es***	-es

Tabla 4: Morfemas de número en los sustantivos

2.4.4.3 Morfemas Facultativos.

Ciertos **sufijos apreciativos** pueden utilizarse con los adjetivos. En muchos casos, no tienen este valor apreciativo, sino que se comportan como sufijos **derivativos**, modificando el lexema del adjetivo.

Se dividen en:

- **Aumentativos**: **azul/*azulón*.**
- **Diminutivos**: En la mayor parte de los casos, el diminutivo tiene un valor afectivo: expresamos con ellos cariño y estimación. Ej. **chico/*chiquito*, revoltoso/*revoltosillo*.**
- **Despectivos**: **pardo/*parduzco*.**

2.4.5 Los Determinantes

Los determinantes son una **clase heterogénea** de palabras. Esto quiere decir que no todas poseen las mismas características. Sin embargo se engloban en una sola clase porque todas cumplen la **misma función** sintáctica: son determinantes.

2.4.5.1 Función.

Hemos visto con anterioridad que los sustantivos nos valen para designar a todos los seres de una especie. Con ***perro*** nos podemos referir a todos los perros. Sin embargo, cuando nosotros utilizamos esa palabra, necesitamos precisar, determinar su significado. Para ello usamos los determinantes. Entonces diremos:

Ese *perro es el mío /* ***Nuestro*** *perro se llama Tobi /* ***El*** *perro* ***del*** *guardia es muy grande /* ***Algunos*** *perros tienen pulgas /...*

Como vemos, la función de los determinantes consiste en precisar el significado del sustantivo al que acompaña. Por ello, tiene que ver también con los adjetivos. De hecho, muchos determinantes reciben también el nombre de **adjetivos determinativos**. Frente a los adjetivos calificativos, que expresan cualidades, los adjetivos determinativos determinan al nombre, señalando su número, orden, pertenencia, situación, etc. Además se diferencian porque los adjetivos calificativos constituyen una clase abierta de palabras, mientras que todos los determinantes son una **clase cerrada.**

2.4.5.2 Clasificación

2.4.5.2.1 Actualización vacía

Podemos utilizar los sustantivos en el discurso sin determinantes, puesto que no siempre son necesarios para su actualización. En estos casos atendemos sólo a su valor **esencial**, es decir, a las notas que los definen como tales, no a unos objetos determinados.

2.4.5.2.2 El artículo

El artículo tiene la misión de introducir al sustantivo en la oración, sin añadir ninguna determinación en concreto; por ello es que **no poseen sustancia semántica**. No tiene ningún tipo de significado.

Hay dos clases de artículos: **indeterminado** y **determinado**.

	Determinado		Indeterminado	
	Masculino	Femenino	Masculino	Femenino
Singular	El	La	Un	Una
Plural	Los	Las	Unos	Unas

Tabla 5: Artículos indeterminados y determinados según su número

El artículo indeterminado sirve como primer presentador: introduce en la oración un sustantivo que no es conocido por el hablante, bien porque sea la primera vez que aparece en la oración, bien porque no esté presente en el contexto comunicativo. Ej. *Ha llegado* ***un*** *mensajero a la oficina y ha traído* ***una*** *carta.*

El artículo determinado sirve para introducir en la oración los sustantivos que ya estaban presentes en las oraciones anteriores o que ya conocíamos de antemano porque estaban presentes en el contexto. Así en la oración anterior aparece l***a oficina***, porque se supone que en ese lugar es donde se ha producido ese mensaje. Luego, podríamos decir: *Ha llegado un mensajero* ***a*** *la oficina y ha traído una carta.* ***El*** *mensajero ha dejado* ***la*** *carta en* ***la*** *mesa y se ha marchado.*

Los artículos **concuerdan en género y número** con el sustantivo, al que siempre acompañan. Sólo hay una excepción. Cuando hay un nombre femenino que comienza por a tónica, el artículo determinado que se emplea es el, para evitar la cacofonía. Ej. *Se han encontrado* ***el aula cerrada****.* Como se verá, en estos casos no se guarda la concordancia entre todos los elementos que dependen del sustantivo.

2.4.5.2.3 Los adjetivos determinativos.

Son aquellos determinantes que poseen cierto significado. Todos ellos pueden realizar otra función que es la de sustituir al sustantivo. En estos casos su significación es distinta: son **pronombres**. Distinguiremos a unos de otros porque los determinantes siempre acompañan al sustantivo, con el que concuerdan en género y número. El pronombre nunca acompaña al sustantivo, puesto que equivale a éste.

Los adjetivos determinativos sirven para determinar la extensión semántica del nombre, añadiendo algún significado gramatical nuevo.

Los adjetivos demostrativos.

Los demostrativos pueden tener valor deíctico o fórico.

- Tienen **valor deíctico** cuando se utilizan para situar el sustantivo que acompañan en relación con elementos del espacio comunicativo (espacio, tiempo o participantes):

Demostrativos				
Formas		Proximidad al hablante	Distancia media	Lejanía del hablante
Singular	Masculino	este	ese	aquel
	Femenino	esta	esa	aquella
	Neutro	esto*	eso*	aquello*
Plural	Masculino	estos	esos	aquellos
	Femenino	estas	esas	aquellas

Tabla 6: Adjetivos demostrativos

(Los señalados con asteriscos sólo son **pronombres**)

- Tienen **valor anafórico** o **catafórico** cuando relacionan el sustantivo al que acompañan con otro elemento mencionado en el texto:

  - **Valor anafórico**: Se refieren a un elemento (SN, proposición, oración) mencionado con anterioridad. Ej. ***Estudias mucho*** *y* ***eso*** *está muy bien.*

  - **Valor catafórico**: Anticipan un elemento del discurso. Ej. *Nos dijo* ***esto****:* ***haced lo que queráis***

Los adjetivos posesivos.

Los posesivos señalan a quién pertenece el objeto designado por el nombre. A veces simplemente establecen una relación imprecisa, sin que se trate exclusivamente de pertenencia. Están referidos a las tres personas gramaticales, pero no expresan persona. Pueden ir **antepuestos** (átonos) o **pospuestos** (tónicos) al sustantivo. Presentan varias formas según su posición:

Posesivos			
		Un poseedor	Varios poseedores
Referido a 1.ª persona	Formas tónicas	mío, mía, míos, mías,	nuestro, nuestra, nuestros, nuestros
	Formas átonas	mi, mis	nuestro, nuestra, nuestros, nuestros
Referido a 2.ª persona	Formas tónicas	tuyo, tuyo, tuyos, tuyas,	vuestro, vuestra, vuestros, vuestras
	Formas átonas	tu, tus	vuestro, vuestra, vuestros, vuestras
Referido a 3.ª persona	Formas tónicas	suyo, suya, suyos, suyas	
	Formas átonas	Su, sus	

Tabla 7: Adjetivos posesivos

Los adjetivos numerales.

Los adjetivos numerales se dividen en **cardinales, ordinales, múltiplos y partitivos**.

- Los numerales **cardinales** expresan una cantidad exacta y preceden al sustantivo. Ej. ***Tres*** *famosos cantantes actuarán en Madrid este año.* La ortografía de los numerales es bastante precisa: Se escriben en una sola palabra hasta *treinta*; a partir de ahí se forman por **coordinación** o ***yuxtaposición*** de palabras (excepción hecha de los que expresan decenas y centenas). Ej. **treinta y dos**, ***doscientos catorce.***

- Los numerales **ordinales** señalan el orden que ocupa el nombre dentro de una serie. Ej. *El* ***Quinto*** *Centenario se celebró con fastuosidad.*

  La serie de los numerales es: primer(o), segundo, tercero, cuarto, quinto, sexto, séptimo, octavo, noveno, décimo, undécimo, duodécimo, décimo tercero,..., vigésimo, vigésimo primero,..., trigésimo, cuadragésimo, quincuagésimo, sexagésimo, septuagésimo, octogésimo, nonagésimo, centésimo.

  Sus formas son, a veces, muy complejas, por lo que es muy frecuente hacer un mal uso de los mismos. Es especialmente frecuente su sustitución por los

partitivos. Sería un rasgo de pulcritud lingüística la correcta utilización de los mismos.

- Los numerales **partitivos** indican fracciones. La única forma propia es **medio**. Los demás se forman con el **cardinal** correspondiente **más** el sufijo **-avo.** Ej. **dieciseisavo**.
- Los **múltiplos** indican multiplicación. Son usuales **doble, triple, cuádruple**. El resto son infrecuentes y se sustituyen por otras fórmulas expresivas.

Los adjetivos indefinidos.

Presentan una débil caracterización semántica. No se puede señalar que todos expresen una **cantidad imprecisa (*mucho, poco, bastante, demasiado*)**. Muchos sólo son **identificadores (*mismo, otro, propio, tal*)**; otros afirman o niegan la **existencia** de algo. (***algún, ningún***). Pueden funcionar como **pronombres**.

Un elemento diferenciador de los indefinidos es su posibilidad de funcionar como adverbios de cantidad, permaneciendo, en estos casos, **invariables**. Se reconocen porque no determinan a un sustantivo, ni lo sustituyen (Pronombres indefinidos), sino que modifican a un adjetivo, a un adverbio, o son complementos circunstanciales del verbo. Ej.

Adjetivo indefinido	Pronombre indefinido	Adverbio-Modificador del adjetivo	Adverbio-Complemento circunstancial del verbo	Adverbio-Modificador de otro adverbio
Tiene **más** coches	No compres **más**	Es **más** alto	¡No fumes **más**!	¡Eso está **más** lejos
Hay **demasiados** coches	Son demasiados	Son **demasiado** contaminantes	Fuma demasiado.	Es **demasiado** pronto

Tabla 8: Adjetivos indefinidos

Los adjetivos interrogativos y exclamativos.

Las formas son comunes: **qué, cuál/-es, cuán (to)/-a/-os/-as.**

Los adjetivos interrogativos preguntan por algo concreto. Ej. *¿**Qué** vestido te pondrás hoy?*

Los adjetivos exclamativos señalan una exclamación ante el sustantivo. Ej. *¡**Qué** alboroto produjo el 92 en España!*

Los adjetivos distributivos.

Indican la forma de distribuirse los sustantivos. Generalmente sólo se utilizan dos: **ambos/-as y sendos/-as.**

Ambos señala a dos individuos de un especie, pero diferenciándolos. Ej. *Los dos porteros se insultaron; el árbitro expulsó a **ambos** jugadores.*

Sendos, sendas significa uno para cada uno: Ej. *A mi hermano y a mí nos han regalado **sendas** bicicletas.*

2.4.6 Los Verbos

Es la parte de la oración que posee más morfemas, aunque estos se presenten en muchos casos **amalgamados**, es decir, unidos. Es lo frecuente en los casos de **persona - número** y **tiempo - modo.** Además de estos morfemas ya señalados, el verbo posee los morfemas de **aspecto y voz.**

2.4.6.1 Base significativa.

El verbo expresa los **"accidentes"** que le ocurren al sujeto. Es la forma que tenemos de contemplar la realidad con respecto al **tiempo**. De otra forma, podríamos decir que el verbo **expresa acciones, procesos o estados respecto a un sujeto, situándolas siempre en el tiempo**.

2.4.6.2 Morfemas del verbo.

El verbo posee los siguientes morfemas: **persona, número, tiempo, aspecto, modo y voz**. No todos son característicos suyo, pues comparte la persona con el pronombre, y el número con los sustantivos y pronombres. El resto de los morfemas son propiamente verbales.

2.4.6.2.1 El tiempo.

El verbo es la parte de la oración que más expresa el tiempo. Existen, lingüísticamente, tres tiempos: **pretérito o pasado, presente y futuro**. Ahora bien, para estudiar estos tres tiempos verbales, debemos considerar que el tiempo, como concepto, es algo absolutamente relativo; depende de la forma en que los hablantes viven esa idea vaga que entendemos por tiempo.

- **El pasado** es, en realidad, el único tiempo que existe, el único del que podemos hablar con seguridad, puesto que ya ha sido vivido. De hecho, hay muchos más tiempos verbales referidos al pasado que al presente o al futuro.
- **El presente** no existe como tal tiempo. En cuanto que no se ha cumplido es futuro; en cuanto se cumple pasa a ser pretérito. Por ello, hay que tener en cuenta que se vive de forma distinta según los hablantes. El presente lingüístico puede tener una duración momentánea o puede abarcar todo nuestro siglo. Ej. *Ahora* ***hace*** *mucho frío. Hoy* ***hace*** *mucho frío.* ***Estamos*** *ahora a plena luz del día.* ***Estamos*** *en el siglo XX.*
- **El futuro** es el tiempo más incierto. Por ello también es el menos utilizado. En muchos casos se sustituye por otros tiempos que expresan mayor certeza.

2.4.6.2.2 El modo.

El modo expresa la **intención del hablante** al comunicar una acción.

- **Modo indicativo.** El hablante puede contemplar la acción de una forma objetiva, real, pensando que el hecho ha sucedido, sucede o sucederá. Entonces estamos ante **el modo indicativo**. Ej. *Detrás de los cristales **llueve**. Hoy **ha llovido** muchísimo. Mañana **lloverá** mucho sobre la zona centro.*

- **Modo subjuntivo.** Pero también puede el hablante expresar su subjetividad sobre la acción del verbo. En este caso, se expresan las acciones como posibles, dudosas, irreales, etc. Es **el modo subjuntivo**. Por ello, las diferencias temporales en el subjuntivo se borran y aparecen menos claras que en el modo indicativo. Además el subjuntivo depende en muchos casos de que el verbo esté o no **subordinado** a otro. Ej. ¡Ojalá llueva hoy! ¡Ojalá llueva mañana! Si nevara mañana, no vendría al instituto.

- **Modo imperativo.** Trataremos en este apartado el imperativo, aunque no está claro que sea un modo de contemplar la realidad. Posee unas peculiaridades que lo distinguen de los otros modos:

  • **Sólo posee una persona, la segunda**. (En singular y plural)

  • **No se puede utilizar con negación**. Hay que recurrir a otras formas verbales para expresar el mandato negativo en español. Ej. Cerrad la ventana / * No cerrad la ventana/ No cerréis la ventana

  • **Sólo cumple la función conativa o apelativa** de la lengua.

  Por todo ello, es un elemento verbal extraño, de una eficacia expresiva grande.

2.4.6.2.3 La persona y el número.

No son dos morfemas característicos del verbo. **Dependen del sujeto, que impone el número y la persona al verbo**. Por ello decimos que el sujeto y el verbo concuerdan en número y persona. Recordamos que sólo hay tres personas:

- Primera: el que habla.
- Segunda: el que escucha.
- Tercera: aquel/la o aquello de lo que se habla.

Advertimos que los sustantivos, al ser siempre elementos de los que se hablan, serán siempre tercera persona. La concordancia en número y persona es el elemento más importante para establecer cuál es el sujeto de una oración.

2.4.6.2.4 Las formas no personales del verbo.

Como acabamos de señalar, la persona y el número no son morfemas característicos del verbo. Por ello hay formas verbales que no expresan persona ni número. Las formas no personales del verbo son **el infinitivo, el gerundio y el participio**. Se caracterizan porque pueden funcionar siempre como verbos; además están emparentados con otras clases de palabras:

- El infinitivo se relaciona con el sustantivo.
- El gerundio se relaciona con el adverbio.
- El infinitivo se relaciona con el adjetivo.

Las formas no personales sí expresan tiempo y, en parte, voz. El infinitivo tiende al futuro; el gerundio, al presente; el participio, al pasado. Los dos primeros pueden conjugarse en activa y pasiva, no así el participio, cuyas formas son pasivas. Todos ellos son muy importantes en la formación de las perífrasis verbales.

2.4.7 Los Adverbios

El adverbio es una clase **heterogénea** de palabras, con una función sintáctica predominante (complemento circunstancial), pero no exclusiva, puesto que puede cumplir otras funciones. Sin lugar a dudas, el adverbio tiene una relación mayor con el verbo, de donde toma el nombre: « **ad-verbum** » = « **junto al verbo** ».

2.4.7.1 Características morfológicas.

Formalmente el adverbio se caracteriza por ser una **parte invariable de la oración**, es decir, es una clase de palabras que no posee género, número, persona, tiempo, etc. Permanece "siempre" con la misma forma; esto es de gran ayuda para distinguir, por ejemplo, los adverbios de cantidad de los adjetivos indefinidos.
No obstante lo dicho, algunos adverbios si que pueden recibir ciertos morfemas:

Apreciativos, como los sustantivos y adjetivos, especialmente los **diminutivos**. No siempre tienen un valor diminutivo; frecuentemente añaden un valor afectivo. Ej. *Ayer me levanté **tempran**ito. **Ahor**ita lo hago.*

Morfemas de grado, como los adjetivos: Ej. *Vive **lejísimos.***

2.4.7.2 Clasificación según sus monemas.

Según su composición morfológica, es decir, según los monemas que las compongan, se dividen en tres clases:

- **Simples**: son aquellos que están compuestos por un sólo monema, con significación léxica. Ej. ***hoy, mañana, tarde, ahora, ahí.***
- **Compuestos**: se forman con un **adjetivo** en grado positivo más el morfema -*mente*, es decir, están compuestas por dos monemas. Por ejemplo: ***buena**mente, **rápida**mente, **feliz**mente.*
- **Locuciones adverbiales**: están compuestas por varias palabras. Forman un conjunto que no es susceptible de ser analizado sintácticamente. Poseen un significado distinto a la suma del significado de las palabras aisladas. Ej. ***a lo loco, a ciegas, a pies juntillas, en un abrir y cerrar de ojos.***

2.4.7.3 Clasificación según su significado.

También cabe clasificarlos según el modo de significar dentro de la oración: deíctica.

Situacionales: Tienen significación deíctica y sitúan la acción respecto al tiempo y lugar dentro del discurso. Por ello, algunos se relacionan estrechamente con los adjetivos demostrativos y pronombres personales, con los que comparten su valor deíctico.

- **Lugar**: aquí, acá, ahí, allí, encima, debajo, cerca, lejos, enfrente, alrededor.
- **Tiempo**: ahora, hoy, ayer, mañana, pasado mañana, pronto, tarde, anoche, antes, después, últimamente, próximamente.

Conceptuales: Tienen una significación permanente, independiente del discurso.

- **Cantidad**: muy (mucho), más, poco, bastante, demasiado, nada...(Se confunden fácilmente con los indefinidos)
- **Modo**: bien, mal, despacio, aprisa, apenas, aposta, así, libremente, cortésmente, a hurtadillas...

Oracionales: Su significado afecta a toda la oración.

- **Afirmación**: sí, claro, ciertamente, en efecto, efectivamente,
- **Negación**: no, nunca, jamás, tampoco
- **Duda**: quizá, tal vez, acaso, posiblemente, a lo mejor, seguramente...

2.4.7.3.1 Los pronombres relativo-adverbiales.

Hay un número determinado de adverbios (**donde, como, cuando**) que están emparentados con los pronombres relativos. Sirven de nexo de proposiciones adjetivas. Tienen un **antecedente**, con el que están relacionados, al que sustituyen en la oración. Ej. *Desde mi ventana se ve el **jardín** **donde** juegan los niños. (CCL)*

La única diferencia entre los pronombres relativos y estos pronombres relativo-adverbiales es que estos **cumplen siempre la misma función** sintáctica: son siempre **complementos circunstanciales de lugar, tiempo y modo**.

2.4.7.3.2 Adverbios interrogativo-exclamativos.

Introducen oraciones interrogativas o exclamativas. Son **cómo, cuándo, dónde, por qué.**

2.4.8 Los Pronombres

Constituyen, al igual que los determinantes una clase **heterogénea** de palabras. Es una clase **cerrada** de palabras, pues tienen un número limitado y fijo de elementos. Como ya hemos señalado, su forma coincide, en muchos casos, con los determinantes.

2.4.8.1 Significación de los Pronombres.

Los pronombres tienen en común **la forma de significar**; con ello se quiere decir no que todos signifiquen lo mismo, sino que se comportan de la misma forma en su manera de significar. Siempre tienen un **significado ocasional**, que depende del **contexto** comunicativo. Ej.

- *¿**Quién** ha entendido **esto**?*
- ***Yo***
- *¿**Tú**?*
- *Sí, **yo***

En cada uno de estos ejemplos, los pronombres cambiarán de significado, dependiendo de quien lo haya enunciado o de cuál sea la situación comunicativa. Ahora bien, su significado gramatical nunca lo pierden; esto es, ***quién*** es siempre pronombre interrogativo (singular); ***esto,*** demostrativo (neutro); **yo**, personal (1ª singular sujeto); **tú**, personal (1ª singular sujeto).

2.4.8.2 Clasificación.

La clasificación de los pronombres coincide, en gran medida, con los adjetivos determinativos

2.4.8.2.1 Los Pronombres Personales.

Significado gramatical.

Tienen un significado gramatical muy preciso: expresan las **tres personas gramaticales:**

- **1ª persona**: el que habla o emisor
- **2ª persona**: el que escucha o receptor
- **3ª persona**: aquel o aquello de lo que se habla.

No es cierto en todos los casos la afirmación de que el pronombre sustituye al nombre. Esto sólo es válido para la tercera persona, pues **yo** y **tú** no sustituyen a ningún elemento en la oración.

Morfemas constituyentes.

Los pronombres personales tienen los morfemas de **género** y **número** (No en todos las personas, como puede verificarse en el esquema). Pero **su morfema característico es el caso**, que se expresa mediante formas heredadas del latín. Esto quiere decir que **la forma del pronombre personal expresa la función sintáctica que desempeña en la oración**, aunque esta diferenciación sólo se cumple estrictamente en la tercera persona.

		Sujeto (y atributo)	Complemento directo	Complemento indirecto	Complemento preposicional
1ªpersona	Singular	yo	me		mí, conmigo
	Plural	nosotros (-as)	nos		nosotros (-as)
2ªpersona	Singular	tú	te		ti, contigo
	Plural	vosotros (-as)	os		vosotros (-as)
3ªpersona	Singular	él,ella, ello	lo, la	le, se	él,ella, ello
	Plural	ellos (-as)	los, las	les, se	ellos (-as)

Tabla 9: Los morfemas constituyentes en los pronombres personales.

2.4.8.2.2 Los Pronombres Demostrativos, Posesivos, Numerales e Indefinidos.

Coinciden en su mayoría con los adjetivos. En todos los casos podemos decir que el pronombre sustituye a un nombre. La diferencia entre unos y otros radica en su relación con los sustantivos: los adjetivos determinativos, como determinantes que son, acompañan siempre al nombre; los pronombres, puesto que sustituyen al nombre, nunca le pueden acompañar.

Además, otra diferencia estriba en que los determinantes no tienen género neutro, pues no hay sustantivos neutros en español; **los pronombres sí tienen género neutro**; en estos casos, el neutro tiene un carácter general y significa un conjunto de elementos. Los posesivos y los numerales ordinales frecuentemente se unen al artículo cuando funcionan como pronombres: *el mío, la tuya, los suyos, la segunda, los cuartos.* Los pronombres indefinidos presentan algunas formas características: sólo son pronombres **alguien, nadie** (tienen un referente), **nada, algo** (tienen un referente no humano).

2.4.8.2.3 Los Pronombres Relativos.

Son pronombres porque sustituyen a un nombre: evitan repetir ese sustantivo. Se llaman relativos porque están relacionados con un sustantivo citado anteriormente en la oración. El sustantivo con el que se relacionan se llama **antecedente**. No sólo son pronombres. Todos ellos son, a la vez, elementos de relación o **nexos introductorios de proposiciones adjetivas o de relativo,** dentro de la cual **cumplen una determinada función sintáctica,** puesto que equivalen al Sintagma Nominal al que sustituyen.

Formas de los pronombres relativos.

Los pronombres relativos son cuatro: **que, cual, quien, cuyo**. Se pueden confundir con los determinantes y pronombres interrogativo-exclamativos. Se diferencian porque los relativos siempre se relacionan con su antecedente, mientras que los interrogativo-exclamativos mantienen una relación distinta con el sustantivo. Además, los pronombres relativos son **átonos**, mientras que los interrogativo-exclamativos son siempre tónicos. Veamos los cuatro que existen en castellano:

- **QUE**: Su forma es siempre la misma, sea cual sea el género y número de su **antecedente**:

  Vimos en la playa a un niño **que** jugaba con la arena (SUJ)

  Vimos en la playa a una niña **que** jugaba con la arena (SUJ)

  Vimos en la playa a unos niños **que** jugaban con la arena (SUJ)

  Vimos en la playa a unas niñas **que** jugaban con la arena. (SUJ)

  Aunque el antecedente es distinto en género y número, el relativo no cambia. Se puede confundir con la **conjunción que**. Para distinguirlo basta con cambiar el relativo que con otro de los relativos que veremos a continuación; si el cambio es posible, sabremos que en ese caso se trata del pronombre relativo **que**.

- **CUAL**: Se utiliza siempre con el artículo. Sus formas son las siguientes: **el cual, la cual, las cuales, los cuales**. Es el relativo más apropiado cuando entre el antecedente y el relativo se interpone una preposición, Ej. *La ventana, junto a **la cual** me siento, tiene rotos los cristales. (CCL)*

- **QUIEN**: Sólo tiene dos formas, en singular y plural: **quien, quienes**. Valen para femenino y masculino. Su única peculiaridad es que su antecedente tiene que tener el rasgo humano. Ej. *El profesor de **quien** te hablé me ha suspendido. (C. Régimen o Suplemento)*

- **CUYO**: Tiene todos los morfemas de género y número. Sus forma son: **cuyo, cuya, cuyos, cuyas.**

  Es el que menos se usa hoy, y en el que se cometen mayores incorrecciones en su uso. Posee varias características que lo hacen especial:

  - **es pronombre relativo**, porque se relaciona con un sustantivo anterior o **antecedente**, con respecto al cual expresa posesión.
  - **es determinante posesivo**, porque expresa la posesión de un objeto o **consecuente** que pertenece al antecedente. Concuerda, como determinante

que es, en género y número con este consecuente. Ej. *Aquel **señor**, **cuyo** perro ves, tiene malas pulgas. (Det.)*

Los pronombres relativo-adverbiales.

Hay un número determinado de adverbios (**donde, como, cuando**) que están emparentados con los pronombres relativos. Tienen un **antecedente**, con el que están relacionados, al que sustituyen en la oración. Ej. *Desde mi ventana se ve el jardín **donde** juegan los niños. (CCL)*

La única diferencia entre los pronombres relativos y estos pronombres relativo-adverbiales es que estos **cumplen siempre la misma función** sintáctica: son siempre **complementos circunstanciales de lugar, tiempo y modo**.

2.6.8.2.4 Los Pronombres Interrogativo-Exclamativos.

Su forma coincide, en parte, con los pronombres relativos. Se diferencian porque los interrogativo-exclamativos, sean pronombres o determinantes, son siempre tónicos (son una clase de palabras que siempre se acentúan); por ello llevan siempre tilde, para diferenciarlos de los relativos. Además, nunca llevan antecedente, como los relativos. Otros están relacionados directamente con los adverbios, puesto que preguntan por los Complementos Circunstanciales, función característica de los adverbios. Son pronombres cuando no acompañan a ningún sustantivo.

Relacionados con los pronombres relativos: **qué, quién/-es, cuál/-es.** (El uso de **cúyo**, como pronombre interrogativo se ha perdido actualmente, pero no era infrecuente hasta los Siglos de Oro)

Relacionados con los adverbios: cuán (to) / -a/ -os/ -as, cómo, cuándo, dónde, por qué

2.4.9 Las Conjunciones

Junto a las preposiciones, son **elementos relacionantes** de la oración. Las conjunciones se pueden clasificar según los elementos que ponen en relación y según sea el tipo de

relación que se establece entre esos elementos. Así, veremos tres tipos de conjunciones, que estudiaremos más detenidamente cuando hagamos el estudio de las oraciones compuestas y complejas.

2.4.9.1 Conjunciones coordinantes.

Unen elementos funcionalmente equivalentes, es decir, elementos con la misma función, sean sustantivos, adjetivos, sujetos, complementos circunstanciales, «proposiciones», etc. En el caso de que unan «proposiciones», no se establece entre ellas una relación especialmente significativa.

- **Copulativas**: y (e), ni
- **Adversativas**: mas, pero, aunque, sin embargo, sino, no obstante, empero.
- **Distributivas**: bien... bien..., ya...ya..., sea...sea..., o...o... (u).
- **Disyuntivas:** o (u).
- **Explicativas**: o (u), esto es, es decir, o sea.

2.6.9.2 Conjunciones completivas.

Son las que nos sirven para unir los diferentes elementos que componen las oraciones complejas. Atención: este término tiene un significado que no coincide con la mayoría de los libros de texto. Será visto con detenimiento más adelante. Oraciones complejas son aquellas en que existe una proposición que depende de un elemento de la oración en que se integran. (Debemos recordar aquí que los pronombres relativos y los relativo-adverbiales introducen proposiciones subordinadas adjetivas y circunstanciales de lugar, tiempo y modo).

Introducen una *proposición subordinada sustantiva*. Las conjunciones completivas son muy pocas:

- Si introducen una proposición enunciativa, la conjunción es **(el) que**. Ej. Elías dice **que irá de vacaciones a México**. **El que venga tu hermano** no me preocupa.

- Si introduce una proposición interrogativa total, la conjunción es **si**. Ej. Elías preguntó **si irían de vacaciones a México**. **Si apruebas o suspendes** me importa sobremanera.

2.4.9.4 Conjunciones subordinantes.

Son las que utilizamos para relacionar estrechamente dos o más oraciones simples. Con ellas se expresan **relaciones lógicas**, como la condición, la causa, la consecuencia, la concesión, la comparación y la finalidad. Habrá, pues, conjunciones condicionales, causales, consecutivas, concesivas, comparativas y finales. Veremos sólo las más usuales, teniendo en cuenta que hay que distinguir entre conjunciones (una sola palabra) y locuciones conjuntivas (dos o más palabras).

- **Condicionales**: si, a condición de que, con tal de que, como.
- **Causales**: porque, pues, como, puesto que, dado que, pues que, ya que.
- **Consecutivas**: tan, tal, tanto...que; luego, conque, así pues.
- **Concesivas**: aunque, a pesar de que, aun cuando, si bien, etc.
- **Comparativas**: más... que, tan... como, menos... que.
- **Finales**: para que, a que, a fin de que, con objeto de, con la intención de que, etc.

Conclusiones

En este capítulo se ha realizado una descripción de las principales características de la morfología de nuestra lengua materna. Se expusieron los modelos lingüísticos empleados tradicionalmente para la descripción de la morfología y finalmente se describieron las principales clases de palabras desde el punto de vista morfosintáctico.

Capítulo 3.

La Morfología desde el punto de vista Computacional

Introducción

Un analizador morfológico automático es una herramienta básica para cualquier sistema de procesamiento de lenguaje natural (PLN). Aunque en algunos sistemas se venía prescindiendo del tratamiento morfológico, sobre todo porque el idioma a tratar era el inglés, usándose diccionarios completos de formas también llamados léxicos desplegados, su uso se ha hecho prácticamente universal como base a cualquier otro tratamiento lingüístico automatizado.

Hasta la fecha se han desarrollado varios sistemas que abordan la morfología. En este capítulo nos proponemos presentarles algunos de ellos y profundizar un poco en la morfología de dos niveles como formalismo más aceptado.

3.1 Consideraciones Generales

Para evaluar la realización y el alcance de un Analizador Morfológico exponemos a continuación los criterios extraídos del trabajo: "Morfología de Estados Finitos" [Alegría].

Poder Expresivo del Formalismo o Modelo Propuesto

Conjunto de fenómenos que pueden ser expresados o analizados por el modelo. En este aspecto se incluye la posibilidad de realizar síntesis y generación en contraposición con sólo permitir uno de los dos procesos.

Forma de Abordar la Morfología

Influenciados principalmente por las teorías lingüísticas y el modelo computacional en que se apoya el formalismo se suelen distinguir dos aproximaciones, una basada en léxicos, en donde raíces y afijos (morfemas) son las unidades básicas y los elementos que gobiernan el proceso y otra basada en paradigma, donde los paradigmas son la base del sistema y el resto de los elementos están en función de ellos.

Forma de Resolución de la Morfotáctica

Forma de especificar las relaciones posibles entre morfemas. Se suelen dividir en dos grandes grupos: morfotáctica de estados finitos y mecanismos de unificación. En el primer caso las relaciones entre morfemas pueden ser vistas como un grafo, siendo los morfemas los nodos, y los encadenamientos posibles los arcos. Los mecanismos de unificación se basan en las gramáticas que se suelen utilizar en sintaxis y son más potentes y flexibles que los de estados finitos, aunque computacionalmente son más complejos.

Especificación de los Cambios Fonológicos

Aunque en la bibliografía aparecen varios métodos, sobresalen entre ellos dos: métodos *ad-hoc* por programa que eran habituales hasta hace algunos años, y métodos basados en Traductores de Estados Finitos bastante habituales hoy en día.

Elementos que se Almacenan en el Léxico

Aunque hay sistemas que funcionan sin las raíces, lo habitual es almacenar morfemas (raíces o lemas y afijos) en el léxico. Dentro de este criterio también cabe distinguir el posible almacenamiento de alomorfos, es decir, si se utiliza más de una representación para una misma unidad léxica, y la deformación deformaciones de los morfemas, es decir, almacenamiento en forma no convencional o canónica.

Sobregeneración y Cobertura

La sobregeneración y la Cobertura, más que del modelo, dependen de la descripción particular que se haga del formalismo. Para la primera es fundamental la granulidad

elegida, lo que si está relacionado en cierto modo con el modelo, ya que el nivel posible está limitado por este. La cobertura suele estar más relacionada con la aplicación para la que se desarrolla el procesador, ya que para un verificador ortográfico nos interesa una cobertura estándar, para un etiquetador nos interesaría la máxima cobertura posible.

3.2 Diferentes Acercamientos.

Antes de introducirnos en la descripción de la morfología de dos niveles, formalismo base de las herramientas desarrolladas en este trabajo, describimos algunos trabajos relacionados con la morfología.

3.2.1 Generación automática de familias morfológicas mediante morfología derivativa productiva.

Tomado del Artículo [Vilares et. al. 2001].

Recursos Lingüístico-Computacionales:

- Lexicón donde cada una de las entradas contiene una forma base, su etiqueta y su lema correspondiente.

Seudo-algoritmo:

El objetivo es generar una supuesta familia morfológica a partir de un término base.

Se define una *familia morfológica* como el conjunto de palabras obtenidas a partir de una misma raíz morfológica mediante la aplicación de mecanismos de derivación.

-Sea F una conjunto que lo utilizaremos para denotar la familia de palabras.

-Sea S una pila que mantiene los componentes todavía no procesados de la familia activa.

1- Inicializar F con el término base Ej. $F = \{rojo\}$.

2- Incluir en S los componentes no procesados de F. Para el ejemplo $S = [rojo]$.

3- Mientras S no se vacíe y existan componentes no procesados en F, se realizan las siguientes acciones:

a) Se extrae el lema situado en el tope de la pila *S* y se le aplican los mecanismos de derivación acordes con su categoría gramatical. La validez de las palabras derivadas (en sus respectivas categorías) es contrastada por medio del lexicón: sólo si están presentes en el mismo se consideran válidos. Si un derivado no es válido, se le aplican las condiciones fonológicas para tratar de obtener uno que sí lo sea. En el ejemplo, *rojo* es extraído de la pila (con lo que ésta queda vacía) y mediante la parasíntesis *en- -ecer* se deriva el verbo *enrojecer*, que es identificado como derivado correcto puesto que pertenece al lexicón.

b) Si se ha obtenido un derivado válido:

1- Si el derivado no había sido previamente procesado, se incluye en *F* y se apila en *S* para ser procesado posteriormente. En el ejemplo, *F = {rojo, enrojece}* y *S= [enrojecer]* en este momento.

2- Si el derivado ya había sido procesado previamente y además pertenece a una familia *F' ≠F*, ambas, *F* y *F'*, se refieren a subconjuntos de una misma familia morfológica. En tal caso, todos los lemas de *F'* son asignados a la familia activa *F*. Denominamos a este fenómeno *transitividad derivativa.*

Podemos observar que el algoritmo opta por sobregenerar, es decir, aplica todos los sufijos posibles obteniendo todos los derivados morfológicamente válidos, los cuales son filtrados mediante el lexicón. De este modo se resuelve el problema de la decisión sobre la validez y aceptación del término derivado a través únicamente de la forma léxica y de su etiqueta, sin considerar otros aspectos.

La derivación regresiva se implementa de modo indirecto por la transitividad derivativa: en vez de derivar el sustantivo a partir del verbo, se espera a que el sustantivo sea procesado para obtener el verbo mediante verbalización denominal.

3.2.1.1 Observaciones:

Para este sistema existen casos que limitan su total correctitud:

Palabras con grafía similar, especialmente aquellas muy cortas: *ano* > *anal* y *ana (medida de longitud)* > *anal.*

Monoptongación de diptongos: fuel (carburante) > folía (baile).

Formaciones parasintéticas: plasta > aplastar (aplanar).

Existencia de más de una acepción en el significado de una palabra: *rancho* > *ranchero* pero no *rancho (comida)* > *ranchero.*

Especialización de significados: golpeador (que golpea) < golpe y golpe (de estado) > golpista.

Sentidos figurados: lince (animal) > lincear (advertir lo oculto).

Una consideración adicional es que está diseñado para analizar un conjunto estático de palabras (las contenidas en el lexicón) y que al utilizar del método como reconocedor y segmentador de formas, no tendría en cuenta el dinamismo y productividad de una lengua al no reconocer formas nuevas.

Según los autores, para mejorar la eficiencia del método se podría utilizar información etimológica o semántica para comprobar si la palabra candidata a derivado guarda o no relación con la palabra primitiva. Esto supondría costos notablemente mayores, sin que se pudiese garantizar la desaparición total de errores. Consideremos por ejemplo el caso de palabras que aun manteniendo una relación etimológica, sus significados hayan variado considerablemente a lo largo del tiempo, por ejemplo *Morfeo (dios del sueño) y morfina (analgésico).* En cuanto a la utilización de información semántica, en el caso de palabras con más de una acepción, habría que desambiguar el sentido de cada palabra a partir del contexto en el que ocurre, lo que elevaría considerablemente los costos computacionales.

3.2.2 A Formal Approach to Spanish Morphology. The COES Tools.

Tomado del Artículo [Rodríguez & Carretero 1994].

Este artículo describe el sistema COES, programa que permite el estudio de la morfología del español, en él se describe una interesante formalización de esta.

Teniendo en cuenta características del español tales como las derivaciones de género y número en los adjetivos y sustantivos, las conjugaciones verbales, los pronombres enclíticos etc.; se definió el modelo formal de la morfología que se explicará a continuación.

Sean los siguientes conjuntos:

X → Lexicón, palabras (lemas) que se usarán para generar las formas flexionadas.

L → Lexemas.

M → Morfemas usados conjuntamente con los conjuntos X y L para generar las formas flexionadas.

W → Conjunto de todas las palabras del idioma Español.

Cada entrada del lexicón (X) tendrá asociado predicados que describirán sus categorías morfológicas.

Relación de categorías morfológicas:

Lexemas de adjetivos y sustantivos (NAL).

Lexemas de verbos regulares (RVL).

Lexemas de verbos irregulares (IVL).

Morfemas de verbos regulares (V).

Morfemas de verbos irregulares (W).

Morfemas nominales de: género y número (P) y sólo de número (S).

Morfemas del gerundio y del participio (X para los verbos regulares y Y para los irregulares).

Morfemas pronominales enclíticos (R para los verbos regulares y O para los irregulares).

Morfemas transitivos enclíticos (T para los verbos regulares y Q para los irregulares).

Morfemas que generan adverbios derivados de adjetivos (M).

Cada una de estas categorías es modelada con macrorreglas que reflejan aspectos particulares de la morfología española. Los lexemas son obtenidos a través del método genérico *lex (lema, <lista de morfemas>)* quien aplica cada morfema del segundo argumento al lema y cuando encuentra uno válido devuelve como lexema el resultado de eliminar del lema el morfema.

Ej.

Macrorregla P:

P(pl1, [masc, fem], plural) → S P(pl2, masc, plural) → ES

P(pl4, fem, plural) → AS P(pl3, masc, plural) → CES

P(gen1, masc, sing) → O P(gen2, fem, sing) → A

P(gen3, masc, sing) → E

Luego al aplicar esta macrorregla a *pastor* y *presidente* obtenemos lo siguiente:

lex(pastora, [gen2, pl2, pl4]) → pastor

gen2(pastor) → pastora

pl4(pastor) → pastoras

lex(presidente, [gen2, pl1, pl4]) → president

gen2(president) → presidenta

pl4(president) → presidentas

Las reglas anteriormente explicadas permiten generar el conjunto *W* a partir de los conjuntos *X* y *M*. Una implementación usando las funciones a continuación descritas de un procesador de estados finitos permite reconocer y generar palabras del idioma Español.

La función *islex* extrae lexemas válidos a partir de morfemas y lemas:

$$\forall x \in M, \forall y \in X, \exists islex(x,y) : M \text{ x } X \rightarrow \{0,1\}$$

$$\text{islex(x, y)}=\begin{cases} 1 - - > si(y-x) \in L \\ 0 - - > si(y-x) \notin L \end{cases}$$

La función *ismor* verifica la existencia de correspondencia entre morfemas y palabras para obtener lexemas, esta función es usada para aplicar la regla de reducción *R*.

$$\forall x \in M, \forall y \in W, \exists ismor(x,y) : M \text{ x } W \rightarrow \{0,1\}$$

$$\text{ismor(x, y)}=\begin{cases} 1 - - > si(y-x) \in L,W \\ 0 - - > si(y-x) \notin L,W \end{cases}$$

Una regla de expansión, *E(x, y, z)*, es una aplicación para obtener una flexión de alguna entrada del lexicón. Cada regla es aplicada cuando *islex* un lexema válido usando el primer parámetro (*x*) y la entrada del lexicón (*y*), el resultado de la regla es una nueva palabra que tendrá la estructura *y-x+z*.

$$\forall x,z \in M, \forall y \in X, \exists E(x,y,z) : M \text{ x } X \text{ x } M \rightarrow W$$

$$E(x,y,z) = \begin{cases} y - x + z & si\ islex(x,y) = 1 \\ 0 & en\ otro\ caso \end{cases}$$

Una clase morfológica es definida como un conjunto de reglas presentes en todos los miembros de la clase, formalmente es definida como:

$$\forall y \in X, \forall_i = 1,...n,\ x_i,\ z_i \in M,\ \exists C(y) :\ X \rightarrow W$$

$$C(y) = \bigcup_{i=1}^{n} E_i(x_i, y, z_i) \cup E(0,y,0)\ \text{C}$$

Una regla de reducción, *R(x, y, z)*, es una aplicación para obtener la raíz de una forma flexionada (lema).

$$\forall x,z \in M, \forall y \in W, \exists R(x,y,z) : M \times W \times M \rightarrow X$$

$$R(x,y,z) = \begin{cases} y - x + z & \text{si } ismor(z,y)=1 \\ 0 & \text{en otro caso} \end{cases}$$

3.2.2.1 Observaciones:

Esta formalización tiene algo importante en su contra: el rendimiento. Lo que se puede notar en el hecho de que es intrínseco al modelo la excesiva cantidad de reglas, por ejemplo, 200 para expresar la conjugación de los verbos regulares y 2,700 para los irregulares. Además sus funciones básicas están expresadas en base a productos cartesianos de conjuntos enormes y por tanto el acceso y el tiempo de respuesta deben verse limitados.

No obstante la limitación del rendimiento resulta un buen material de estudio para analizar la morfología del español, por la especificidad con que representa las reglas y, más aún, teniendo en cuenta que es software libre y que, por lo tanto, puede adaptarse a otras necesidades.

3.2.3 A logical approach to the lemmatization of computational lexica

Tomado del Artículo [Mills 1999].

En este artículo se marcan las pautas a seguir para construir un diccionario con base en PROLOG que pueda ser utilizado como tal y que además contenga información morfológica, permitiendo el reconocimiento de formas derivadas y flexionadas, así también como generarlas. En lo consiguiente trataremos de mostrar la estructura del diccionario y deducir la forma en que se realizan los mencionados análisis.

La estructura:

Cada entrada E es un subconjunto del diccionario $D : E \subseteq D$

Siendo ambos conjunto de conjuntos. La entrada consiste en un par ordenado de dos conjuntos <A,B>. El primer conjunto, A, es el lema y B es el resto de la entrada. Esto se puede expresar en PROLOG como sigue:

entry (lemma, Rest).

El propósito del lema es identificar la unidad léxica, para localizarla en el sistema morfológico y describir su forma, la que puede incluir indicaciones sobre la pronunciación. En esencia el lema en la entrada representa un paradigma de flexión y derivación. Todas las unidades léxicas, deben ser registradas en el diccionario como lemas, incluyendo palabras, derivados y compuestos, morfemas de flexivos, afijos, términos multipalabras, frases verbales, palabras de otros idiomas asimilados por la lengua, proverbios, variantes gráficas, abreviaciones y nombres. La función del lema es posible a partir de cinco elementos: la forma base, pronunciación, categoría gramatical, formas oblicuas (sufijos admisibles por el lexema o forma base) y el desambiguador semántico.

<A,B,c,D,e>

Esto se expresa en prolog de esta forma:

Lemma (Baseforms, Pronunciation, Par-of-spech, Oblique-forms, Semantic-disambiguator).

Si existen variantes ortográficas de la palabra se introducirá como Baseform una sola considerada como primaria a elección del lexicógrafo y se mantiene una referencia a la otra. La Baseform forma un par ordenado <a, B>, donde "a" es un elemento u B es un conjunto con las variantes ortográficas, o referencias. En el diccionario de PROLOG, la forma base (Baseform) es instanciada mediante una lista cuya cabeza es la forma primaria y la cola las variantes:

[Canonical-form | Variants]

Cuando el afijo es altamente productivo no es posible establecer una entrada para cada combinación posible en la que el puede ocurrir, en estos casos se establecen ellos como entradas, considerándolos como un conjunto perteneciente a una categoría gramatical.

La pronunciación, también puede tener más de una variante, por lo que forma un par ordenado semejante al de la forma base:

[Preferred_pronunciation | Variants]

Los otros elementos del lema informan al usuario acerca de las características flexivas, si lo permite, y sobre la clase a la que pertenece. Esto se puede indicar explícitamente por todas las formas del paradigma, o por un identificador que referencie a los paradigmas anexados al diccionario. En un diccionario amplio, es necesario indicar todas las aberraciones de una unidad léxica perteneciente a un paradigma y restringir las flexiones y derivaciones no admisibles.

Finalmente la estructura de cada entrada es la siguiente:

```
entry(
    lemma(
            [Canonical-form | Variants],
            [Pronunciations],
            Part-of-speech,
            [Oblique-forms],
            Semantic-disambiguator
            ),
        rest
        ).
```

La cláusula PROLOG para el lexema "bar" del inglés sería la siguiente:

```
entry(
```

```
lemma(
        bar,
        /bar/,
        m,
        [-rrow],
        botanical
       )
  [branch, (esp. growing)]
  )
```

que indica la entrada base como "bar", la pronunciación /bar/, la categoría gramatical como "m", la forma oblicua "-rrow" para la formación de la palabra derivada "barrow", la indicación semántica de que pertenece a la botánica, y su significado es "branch, (esp. growing)".

Según el ejemplo se puede deducir la necesidad de algún procesamiento para la derivación, como las reglas morfológicas utilizadas por otros trabajos. Esto se evidencia claramente en el hecho de que "barrow" no es el resultado de la simple concatenación de "bar" + "rrow". Desgraciadamente este hecho no se explica en el artículo, ni tampoco la particularidad de establecer los morfemas más productivos como entradas en el diccionario.

Adicionalmente se definen predicados para la extracción de información del diccionario, algunos los mostramos a continuación.

Wordlist→ Retorna una lista de formas canónicas.

Pos_list → Retorna una lista de las categorías gramaticales.

Pos_list(X)→Retorna una lista de formas canónicas cuya categoría gramatical es X.

Search(X,Y)→Retorna todas las entradas para la forma base X y su categoría gramatical Y.

Search(X)→Retorna la entrada la forma base X con su pronunciación, categoría y desambiguador semántico.

Etc.

3.2.3.1 Observaciones:

Uno de los aspectos más significativos de este enfoque es lo inherente al paradigma de programación declarativo, la facilidad para expresar predicados que realicen búsquedas específicas en el diccionario.

Algo importante a señalar es que los aspectos referentes a la morfotáctica no quedan claros y en su lugar se tiende a plantear explícitamente la combinación de lexemas y morfemas, lo que puede hacer incluso intratable computacionalmente todas las unidades léxicas de una lengua tan flexiva y derivativa como el español.

3.2.4 GRAMPAL: A Morphological Processor for Spanish implemented in Prolog.

Tomado del Artículo [Moreno].

Este sistema, como estrategia general, opta por expresar en predicados, todos los posibles alomorfos para los lexemas y morfemas. Esta característica permite que la generación y el reconocimiento se limiten a la concatenación de morfemas como único mecanismo, lo que simplifica drásticamente las reglas.

Todas las entradas en el diccionario se definen como predicados que corresponden a las categorías morfológicas; a continuación hacemos un inventario completo de estos:

"w" → formas completamente flexionadas.

"wl" → sustantivos y adjetivos que pueden aceptar morfemas de número.

"vl" → lexemas verbales.

"nl" → lexemas nominales y de adjetivos.

"vm"→ morfemas verbales.

"ng" → morfemas nominales de género.

"nn" → morfemas nominales de número.

Los argumentos de algunos predicados son los siguientes:

w(Lemma, Category, Pers.Num, Tense.Mood).

w(Lemma, Category, Gender, Number).

wl(Lemma, Category, Number.Type.List, Gender, Number).

vl(Lemma, Category, Conjugation, Stem.Type.List, Suffix.Type.List).

vm(Pers.Num, Tense.Mood, Finiteness, Conjugation.List, Stem.Type, Suffix.Type).

nl(Lemma, Category, Gender.Type.List, Number.Type.List, Gender, Number).

ng(Gender.Type, Gender, Number). nn(Number.Type, Number).

Ejemplos de entradas:

vm(no,part,nofin,[2,3],99,reg) --? [ido].

vm(no,part,nofin,[2,3],99,part1) --? [o].

vl(imprimir,v,3,[100],[reg]) --? [imprim].

vl(imprimir,v,3,[99],[part1]) --? [impres].

Se introducen algunas características contextuales como átomos que imponen restricciones en la concatenación de morfemas en las reglas de unificación. Estas características nunca son filtradas por el nodo padre de la regla en el análisis. Características atómicas multievaluadas son permitidas en el mecanismo de unificación,

siendo interpretadas con la unión de valores atómicos y representadas como una lista. Esta unión se utiliza solo para características contextuales (tipo de raíz, tipo de sufijo, conjugación, género, número) con el objetivo de mejorar su eficiencia.

Para la conjugación verbal, se le asigna un identificador a cada una de las conjugaciones en cada tiempo y modo. Para cada par XY de dígitos, X indica el tiempo y el modo, Y se codifica de la siguiente manera:

1- Prim. pers. del sing.

2-Seg. Pers. Del sing.

3-Ter. Pers. Del sing.

4-Prim. pers. del plur.

5-Seg. Pers. Del plur.

6-Ter. Pers. Del plur.

Luego la tabla sería como sigue:

pres ind 11 12 13 14 15 16

impf ind 21 22 23 24 25 26

indf ind 31 32 33 34 35 36

fut ind 41 42 43 44 45 46

pres subj 51 52 53 54 55 56

impf subj 61 62 63 64 65 66

cond 71 72 73 74 75 76

imper 82 83 85 86

inf 00 ger 90 part 99

Cada una de las 99 formas flexivas es representada por un número y el código adicional 100 es utilizado para indicar la unión de todos ellos (usado para los verbos regulares). La característica contextual para el tipo de raíz (stem type, "stt") es utilizada para identificar la raíz verbal y la terminación correspondiente a cada forma, mientras que la característica contextual para el tipo de sufijo (suffix type "sut") distingue entre el conjunto de alomorfos del morfema flexivo mediante un conjunto de valores como: reg, pres, pret1, pret2, fut, cond, imp, subj, imper, infin, ger, part1, part2. Dos características contextuales similares se establecen para las raíces y morfemas nominales. A continuación mostramos algunos ejemplos:

Morfemas de adjetivos y sustantivos

ng(mas1,masc,sing) --? [o].

ng(mas2,masc,sing) --? [e].

ng(fem,fem,sing) --? [a].

nn(plu1,plu) --? [s].

nn(plu2,plu) --? [es].

Algunas entradas de sustantivos

nl(presidente, n, [mas2,fem], [plu1], ., .) --? [president].

wl(doctor, n, [plu2], masc, sing) --? [doctor]. nl(doctor, n, [fem], [no], masc, sing) --? [doctor].

wl(bambu1, n, [plu1, plu2], masc, sing) --? [bambu1].

Estas entradas permiten el análisis y la generación de las formas: *presidente, presidenta, presidentes y presidentas* para el lema presidente; *doctor, doctora, doctores y doctoras* para doctor y *bambús y bambúes* para bambú.

Las reglas del modelo es realmente pequeño debido a que la mayor parte de la información ya está contenida en el lexicón. En particular las formas verbales flexionadas se analizan o generan mediante dos reglas, siendo realmente una sola, ya que la segunda surge por el uso del valor 100 como un conjunto.

Reglas para la flexión verbal

-w(Lex, Cat, PerNum, TensMood) --? vl(Lex, Cat, Conj, SttL, SutL), vm(PerNum, TensMood, ., ConjL, Stt, Sut), member(Conj,ConjL), member(Stt, SttL), member(Sut,SutL).

-w(Lex, Cat, PerNum, TensMood) --? vl(Lex, Cat, Conj, [100], SutL), vm(PerNum, TensMood, ., ConjL, ., Sut), member(Conj,ConjL), member(Sut,SutL).

La flexión nominal es un poco más complicada debido a la combinación de dos morfemas flexivos, género y número en algunos casos. El modelo necesita 4 reglas para manejar esto. La primera es para el singular, donde la raíz tiene que ser concatenada con un sufijo de género (niñ-o, niñ-a); la segunda es para los plurales donde un sufijo adicional de número es agregado (niño-s); la tercera construye plurales de una raíz alomorfa y un morfema plural (león / leon-es) y la cuarta valida las palabras en singular obtenidas por la primera regla sin concatenación adicional.

Reglas flexivas de sustantivos y adjetivos

wl(Lex, Cat, [plu1], Gen, Num)--? nl(Lex, Cat, GetL, ., ., .), ng(Get, Gen, Num), member(Get, GetL).

w(Lex, Cat, Gen, Num) --? wl(Lex, Cat, NutL, Gen, .), nn(Nut, Num), member(Nut, NutL).

w(Lex, Cat, Gen, Num) --? nl(Lex, Cat, ., NutL, Gen, .), nn(Nut, Num), Nut=plu2, member(Nut, NutL).

w(Lex, Cat, Gen, Num) --? wl(Lex, Cat, ., Gen, Num).

Para su funcionamiento es necesario incluir un segmentador que provea al parser de las posibles segmentaciones de la palabra a partir de los afijos contenidos en el diccionario. El segmentador se construyó como un predicado no determinista que encuentra todas las posibles segmentaciones de la palabra.

3.2.4.1 Observaciones:

Realmente esta es una implementación que logra expresar de forma simple los fenómenos presentes en la morfología flexiva, demostrando las potencialidades de los lenguajes declarativos. El inconveniente está en que al optar por la representación de todos los alomorfos, sobre todo de los lexemas, se priva de la posibilidad de reducir considerablemente el tamaño del lexicón, ya que estos alomorfos frecuentemente son deducibles a partir de reglas, como es el caso de los alomorfos producto de los cambios ortográficos: *pez> pec-es, explic-ar>expliq-ue.*

3.2.4 El tratamiento de la morfología flexiva del castellano mediante reglas de dos niveles en una gramática de unificación.

Tomado del Artículo [Carulla Oosterho]

Este es un analizador morfológico creado en el marco del proyecto europeo LSGRAM (LRE-61029) cuyo objetivo es desarrollar gramáticas de unificación para la mayoría de las lenguas de comunitarias utilizando la plataforma ALEP ofrecida por la misma UE.

El analizador morfológico está orientado al tratamiento de la flexión del español y para esto la referida plataforma ofrece un segmentador de palabras basado en la técnica de dos niveles y un analizador (parser) basado en unificación.

El formalismo de dos niveles de la plataforma ALEP sigue en su concepción básica la propuesta de su creador Kimmo Koskenniemi (1984), con reglas que establecen una proyección entre una cadena de superficie y una cadena léxica (lematizada), ampliándolas con estructuras de rasgo para expresar información léxica incorporando las propuestas formuladas por Bear (1988) y Trost (1990) entre otros.

Ej:

[] [u,e] [] => [] [o] [],

traíz_verbal:{diptongo=sí}.

En este caso el morfema léxico resultado de la proyección de dos niveles debe ser del tipo *traíz_verbal* y contener el rasgo *diptongo=si*.

El formalismo ofrece también la posibilidad de usar diacríticos (caracteres abstractos) en el léxico. Así, por ejemplo, podemos codificar la vocal temática que esté sujeta a variaciones alomórficas mediante un diacrítico, al cual las reglas de dos niveles proyectan todas las posibles realizaciones superficiales (p.ej. la raíz del verbo *poder* como *'pOd'*).

Las características generales para la segmentación son las siguientes:

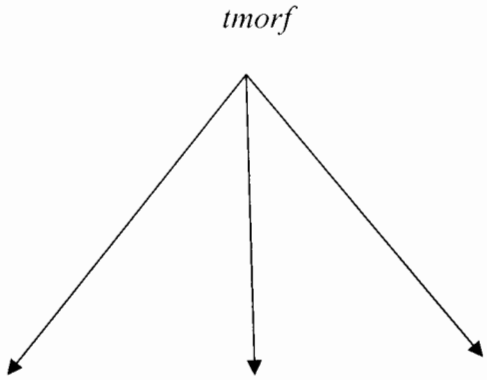
Se distingue entre flexión nominal y la verbal.

Se sigue una estrategia de segmentación en dos unidades: raíz y terminación, sin hacer distinción entre morfemas de número, persona, tiempo, etc.

No se distingue entre terminaciones verbales y nominales para la segmentación, puesto que no se diferencian en la cadena de caracteres sino en la información morfosintáctica.

Codificación léxica

Partiendo de la estrategia de segmentación se tiene una jerarquía de tipos para caracterizar las entradas léxicas, donde *'tmorf'* es un supertipo que contiene como subtipos *'tno_flexivo'* (para las categorías que no flexionan), *'traiz_verbal'* y *'traíz_nominal'*. Las terminaciones son todas del tipo *'tmorf'*.:



Estos tipos llevan los rasgos siguientes:

tmorf

morfema (rasgo atómico que contiene la cadena de caracteres)

ms_key (rasgo atómico que contiene la forma lematizada de morfema)

último = si/no (rasgo que indica la posición que el morfema puede ocupar dentro de la palabra)

eliminación_e=si/no (rasgo que hace referencia a la necesidad de borrar el infijo '-e-' en la formación de plural cuando este no lleva información morfosintáctica (p.ej. ordenador-e-s vs. profesor-es)

tno_flexivo (no contiene rasgos específicos de su tipo)

último= *si*

traíz_verbal

diptongo= si/no

diptongo_red= si/no

cierre = *si/no*

último = *no*

traíz_nominal (no contiene rasgos específicos de su tipo)

último = si/no/_ (la variable anónima '_' es para raíces que no requieren necesariamente un sufijo flexivo)

De la combinación de estos rasgos resultan cuatro grupos de raíces nominales y cinco de raíces verbales.

Reglas basadas principalmente en la información léxica

Para expresar cambios que no responden a necesidades fonológicas u ortográficas, sino que forman parte de las características idiosincráticas de cada raíz verbal como la diptongación y el cierre vocálico, se adiciona a las reglas de dos niveles una estructura de rasgos que mediante la unificación con los rasgos de las entradas léxicas permite su correcta aplicación.

La vocal temática de algunos verbos se diptonga en el presente de indicativo (primera, segunda y tercera persona singular y la tercera persona plural) y en el presente de subjuntivo (primera y segunda persona singular y tercera persona plural) es decir, cuando la raíz va seguida de las terminaciones –o, -as, -a, -an, -e, -es, -en. La descripción exacta del contexto derecho requiere de dos reglas, una para el contexto {o,a,e} seguido de final

de palabra (=) y otra para el contexto {a,e}, {s,n}, }{=}. Para la diptongación de la vocal temática 'o' en 'ue' resulta en las reglas siguientes:

tlm_regla (ue_diptongo1

[] [u,e] [**A**, **B**, =] ⇔ [] ['**O**'] [],

tlmws:{tmorf => traíz_verbal:{diptongo=si}},

A en {b,c,d,g,l,n,ñ,r,v,z}

B en {o,a,e }).

tlm_regla(ue_diptongo2

[] [u,e] [**A**, **B**, ***C***,=] ⇔ [] ['**O**'] [],

tlmws:{tmorf => traíz_verbal:{diptongo=si}},

A en {b,c,d,g,l,n,ñ,r,v,z}

B en {a,e }

C en {s,n}).

La variable A se refiere al conjunto de consonantes con que terminan las raíces verbales. Su especificación contribuye a una mayor restricción en la aplicación de las reglas. Las dos reglas anteriores se aplican cuando la raíz termina en una consonante (*rodar, doler, etc.*)

3.2.4.1 Observaciones:

Las limitaciones importantes de este enfoque, para nuestro objetivo, están en el hecho de no proveer información morfotáctica de forma natural, lo que dificulta el análisis de casos donde la regularización produce segmentaciones idénticas como:

puedo → pOd + o

pudo → *pOd + o*

3.2.6 SEGMORF: un formalismo para analizadores morfológicos de dos niveles.

Tomado del Artículo [Badia et. al.].

Para la morfología de dos niveles presentada por Koskenniemi la morfografemia se establece mediante reglas de dos niveles (RDNs) y la morfotáctica se caracteriza mediante clases de continuación o mediante una gramática de unificación para construir palabras (GP). Separando los dos procesos, se pretende que se puedan expresar de manera natural ambas clases de relaciones. No obstante, como las RDNs intentan encontrar una descomposición léxica de una cadena superficial y como la GP intenta construir una estructura de palabra a partir de aquella descomposición, hay varios fenómenos morfológicos que plantean problemas importantes.

Por un lado, la GP parece requerir información sobre la aplicación de las RDNs, como en los casos (a)-(b) (para seleccionar la descomposición correcta, debería tener información sobre qué regla se ha aplicado).

a) puedo (verbo, 1ra, sing., pres.) => pod (raíz-verbal) + o (1ra, sing., pres.)

b) pudo (verbo, 3ra, sing., pret.) => pod (raíz-verbal) + o (3ra, sing., pret.)

Esto muestra la necesidad de una buena interacción entre las RDNs y las GP. Según [Trost 1990] la forma de tratar esta interacción prevé que:

La información sobre la regla aplicada tiene que ser traspasada a la GP.

Hay que poder restringir la aplicación de algunas RDNs a ciertas clases de morfemas.

En esta propuesta la transferencia de información se concibe como la unificación de la estructura de rasgos asociada a la regla con la estructura de rasgos asociada al morfema léxico. Aunque esta propuesta funciona, presenta algunos problemas, principalmente en relación con el primer requisito:

La GP no está motivada independientemente: a menudo se la concibe como un proceso que valida la aplicación de las RDNs y que filtra las descomposiciones erróneas. Por lo tanto, su formulación depende de las RDNs.

La GP es difícil de manejar, puesto que tiene que adaptarse a la interacción de las distintas reglas que seleccionan su propio contexto.

Para superar las deficiencias anteriores, el tratamiento de los fenómenos morfológicos en que interviene la interacción entre los componentes morfografémico y morfotáctico en el marco de la morfología de dos niveles tiene que cumplir con los siguientes requisitos:

La GP debería ser lo más (modular) independiente posible.

Hay que restringir la aplicación de las RDNs a ciertas clases de morfemas.

Debería ser posible expresar RDNs de manera que el contexto morfológico sea tenido en cuenta junto al morfografémico.

El contexto morfológico:

Dadas las siguientes abreviaturas:

LMC = Contexto morfotáctico izquierdo (Left Morphotactical Context)

RMC = Contexto morfotáctico derecho (Right Morfotactical Context)

Se define el contexto morfotáctico como un par <LMC, RMC>, donde tanto LMC como RMC son una secuencia de cero o más descripciones morfotácticas. Una descripción morfotáctica incluye información morfosintáctica simple, como Categoría Gramatical, concordancia, etc.

Para que la aplicación de una regla sea válida los morfemas ya encontrados deben satisfacer el LMC de la regla, y el resto de los morfemas que se obtendrán posteriormente deben satisfacer el RMC.

Para tratar los casos de "pudo" - "puedo" se necesitan las reglas siguientes:

Rule diptongación:{

[] [ue] [] ⇔ [] [o] []

[] [raíz_verbal + (sufijo_verbal, presente)]

}

Rule cierre: {

[] [u] [] ⇔ [] [o] []

[] [raíz_verbal + (sufijo_verbal, pasado)]

}

Note que son las propias reglas las que seleccionan los morfemas adecuados; la desambiguación se realiza los más pronto posible (de hecho, al consultar el diccionario); la GP no tiene que comprobar qué regla se ha aplicado.

3.2.6.1 Observaciones:

La novedad de este formalismo está en la posibilidad de asociar contextos morfotácticos a las RDNs, característica que no está presente en el componente de dos niveles de la plataforma ALEP.

3.3 Morfología de Dos Niveles

El formalismo más empleado para la representación computacional de la morfología es el conocido como **morfología de dos niveles** propuesta por Kimmo Koskenniemi en 1983, que se enmarca dentro de la morfología de estados finitos, disciplina, en la que se agrupan todas las técnicas empleadas en el tratamiento de la morfología que utilizan como soporte los Autómatas de Estados Finitos. La implementación más conocida y referenciada en la bibliografía es la realizada por Evan L. Antworth en la construcción de PC-KIMMO.

3.3.1 Conceptos Básicos

3.3.1.1 Traductores de Estados Finitos (FST)

Un FST es un autómata que reconoce o genera pares de cadenas, tiene una función más general que un Autómata de Estados Finitos (*FSA*) pues mientras este define un conjunto de cadenas (lenguaje formal) aquel define relaciones entre conjuntos de cadenas.

Un FST puede ser entendido de cuatro maneras diferentes:

Reconocedor: un formalismo que toma pares de cadenas y comprueba que pertenezcan a un determinado lenguaje.

Generador: una máquina que produce pares de cadenas de determinado lenguaje. De manera que la salida es si o no, y el par de cadenas.

Traductor: una máquina que lee cadenas y devuelve otras.

Relacionador de conjuntos: una máquina que establece relaciones entre conjuntos.

Más formalmente, una definición de los FST es la siguiente:

Definición:

Un FST es una 6-tupla $(\Sigma_1, \Sigma_2, Q, i, F, E)$ donde:

Σ_1 es un alfabeto finito que contiene los símbolos de entrada.

Σ_2 es un alfabeto finito que contiene los símbolos de salida.

Q es un conjunto finito de estados.

$i \in Q$ es el estado inicial.

$F \in Q$ es el conjunto de estados finales.

$E \subset Q \times \Sigma_1^* \times \Sigma_2^* \times Q$ es el conjunto de arcos.

Dos propiedades muy útiles de los FST son la **inversión** y la **composición**.

Inversión: La inversión de un Traductor $T(T^{-1})$ simplemente intercambia las etiquetas de entrada y salida. Si $(q_i,a,b,q_j)\in T$ entonces $(q_i,b,a,q_j)\in T^{-1}$.

Composición: Si T_1 mapea desde el lenguaje L_1 a L_2 y T_2 mapea desde L_2 y L_3 entonces $T_1 \circ T_2$ mapea desde el lenguaje L_1 a L_3.

Además de las dos propiedades expuestas anteriormente, todas las que poseen los FSA pueden ser aplicadas a los FST, y en consecuencia los algoritmos si consideramos el alfabeto como un par de elementos compuestos por un elemento del alfabeto de entrada y uno del alfabeto de salida.

Definición:

Si $T=(\Sigma_1,\Sigma_2,Q,i,F,E)$ es un FST, entonces su "Underlying FSA" ó Autómata Finito Subyacente (Haciendo una traducción al español) (Σ,Q,i,F,E') se define como:

$$\Sigma=\Sigma_1\times\Sigma_2$$

$$(q_1,(a,b),q_2)\in E' si (q_1,a,b,q_2)\in E$$

Aunque ya enunciamos anteriormente que los Traductores admiten todas las operaciones definidas para los FSA, para asegurar que la intersección y la composición sean cerradas debe restringirse el conjunto de los FST a los conocidos como "Letter Transducers".

Definición:

Si $T_1=(\Sigma_1,\Sigma_2,Q,i,F,E_1)$ es un FST tal que $\varepsilon\notin|T_1|(\varepsilon)$ entonces existe un $T_2=(\Sigma_1,\Sigma_2,Q_2,i_2,F_2,E_2)$, denominado "Letter Transducer", tal que:

$$|T_1|=|T_2|$$

$$E_2\subset(Q_1\times(\Sigma_1\cup\{\varepsilon\})\times(\Sigma_2\cup\{\varepsilon\})\times Q_2)$$

$$E_2\cap(Q_1\times\{\varepsilon\}\times\{\varepsilon\}\times Q_2)=\phi$$

En el caso de la intersección se añade la restricción de que los traductores deben ser ε-free.

Definición:

Un FST $T = (\Sigma_1, \Sigma_2, Q, i, F, E)$ es llamado ε-free Letter Transducer si $E \subset Q \times \Sigma_1 \times \Sigma_2 \times Q$

En este punto hacemos notar que en la práctica, para suplir la necesidad de equiparar cadenas de longitudes diferentes, lo que se hace es introducir un carácter, comúnmente "0", para al final eliminarlo mediante la composición.

3.3.2 Descripción del Formalismo

Como ya hemos mencionado, en 1983 Koskenniemi definió el modelo computacional para morfología conocido como modelo de dos niveles. Este modelo dio un gran impulso al tratamiento mofo-fonológico y ha tenido gran aceptación debido a las siguientes características:

Es un modelo general, por lo menos a nivel de encadenamiento de morfemas, y aplicable a la mayoría de las lenguas.

Diferencia totalmente el conocimiento lingüístico y el algoritmo.

Es válido tanto para análisis como para generación.

Se diferencian para cada palabra dos niveles o representaciones, el convencional o de *superficie* y el de profundidad o *léxico*. Debido a esta característica se pueden evitar los alomorfos.

Para el tratamiento de los cambios fonológicos, en lugar de utilizar las reglas de reescritura propias de la fonología generativa se emplean reglas paralelas, lo que resulta más sencillo tanto desde el punto de vista conceptual como computacional.

La complejidad computacional no es demasiado elevada, lo que permite la construcción de sistemas reales en microordenadores.

3.3.2.1 El Sistema Léxico

El Sistema Léxico almacena el conjunto de morfemas y la información morfotáctica. Está compuesto por los siguientes tres elementos básicos: entradas léxicas, subléxicos y clases de continuación.

Cada **entrada léxica** se compone de tres campos:

- *Expresión léxica*: es una secuencia de caracteres léxicos. Estos caracteres pueden ser los *convencionales* o de superficie, *morfofonemas* y *marcas de selección* (diacríticos).
- *Clase de Continuación*: sirve para indicar el conjunto de morfemas que pueden encadenarse posteriormente.
- *Información Morfológica*: información que se quiere obtener como resultado del análisis.

Los elementos del léxico están agrupados en subléxicos según sus características morfotácticas, y más concretamente en función del conjunto de morfemas que les pueden preceder. Debido a esto la organización en subléxicos resulta a veces artificial desde el punto de vista lingüístico. La definición de un subléxico consiste en su nombre en su nombre o identificador, sus características y el conjunto de entradas que lo componen. Las características sirven para marcar subléxicos con rasgos morfológicos comunes.

Una *clase de continuación* es un conjunto de subléxicos, que constituye una unidad desde el punto de vista de la morfotáctica y puede ser asimilada a un paradigma. Especifica que cualquiera de los morfemas incluidos en los subléxicos pertenecientes a la clase de continuación pueden ser encadenados inmediatamente después del especificado.

3.3.2.2 Reglas de dos Niveles

La mayor aportación de Koskenniemi fueron las reglas de dos niveles para describir los cambios morfofonológicos por lo que algunos autores prefieren para el formalismo el nombre de fonología de dos niveles.

Las reglas de dos niveles controlan el emparejamiento entre la representación léxica y la de superficie. Las reglas se compilan en FST paralelos y un par de caracteres (léxico-superficie) será aceptado solamente si se acepta en todos los autómatas. No hay ningún tipo de representación intermedia entre los dos niveles mencionados y esta es la diferencia fundamental con la fonología generativa. Durante el análisis se buscan las representaciones léxicas correspondientes al nivel de superficie dado, y durante la generación lo contrario.

La sintaxis de las reglas ha sufrido distintas modificaciones, sobre todo enfocadas a facilitar el trabajo de un compilador automático pero; de forma general el formato es el siguiente:

cp op lc _ rc

Correspondencia (cp) son los pares de caracteres (léxico: superficie) que controlan la regla.

Operador (op), especifica que tipo de relación se establece entre la correspondencia y el contexto que se le asigna. Esta relación puede ser de cuatro tipos: *restricción de contexto* (=>), *coerción de superficie* (<=), composición de ambas (<=>) y *establecimiento de prohibición* (/<=). El significado de cada operador se expresa en la *Tabla 10.*

Contexto (lc _ rc), delimita en que casos se produce el cambio especificado. El carácter "_" separa el contexto a la izquierda (lc) y el contexto a la derecha (rc). En ambas partes del contexto se especifican series de pares de caracteres y se pueden utilizar expresiones regulares para expresar contextos complejos.

op	ejemplo	interpretación
=>	l:s => lc _ rc	el carácter léxico *l* se convierte a nivel de superficie en *s* si el contexto es *lc _ rc.*
<=	l:s <= lc _ rc	En el contexto *lc _ rc l* **siempre** se convierte en *s.*
<=>	l:s < => lc _ rc	El carácter *l* se convierte en *s* **solamente** en el contexto

		lc _ rc.
/<=	l:s /<= lc _ rc	*L* **nunca** se convierte en *s* en el contexto *lc _ rc.*

Tabla 10: Semántica de los operadores empleados en las reglas de dos niveles.

3.3.2.3 Traductores Léxicos

La complejidad computacional básica del modelo de dos niveles viene del uso de reglas on-line. Debido a que un mismo carácter de un nivel se puede emparejar con varios del otro nivel y que puede haber elipsis de caracteres, en muchos momentos se deben seguir varios caminos de análisis por lo que se utilizará *backtracking*. A pesar de que la complejidad estructural viene de las reglas, el léxico, al tener incluida la información morfotáctica, también es causa de perdida de eficiencia. La razón es que una clase de continuación es un conjunto de subléxicos por lo que cuando se llega al final de un morfema se debe seguir por distintos puntos del sistema léxico aunque la mayoría de los caminos no tendrán éxito.

Una mejora espectacular en la eficiencia de estos sistemas es la propuesta de Karttunen de los llamados Lexical Transducer (Traductores Léxicos haciendo la traducción al español) [Karttunen 1994]. En esta propuesta se pueden evitar las representaciones arbitrarias utilizando las correspondientes formas canónicas acompañadas de rasgos morfológicos y además se pueden componer una serie de sistema de reglas permitiendo mayor expresividad, claridad y modularidad en la descripción.

En los traductores léxicos se pueden emparejar en el léxico formas flexionadas con sus correspondientes canónicas:

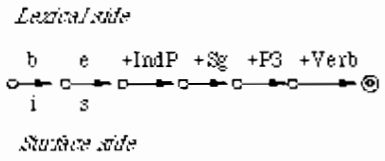


Figura 1: Ejemplo de un camino en el que se establece la correspondencia entre una forma léxica y una de superficie y además se codifica información léxica.

La forma estándar de construcción de un Traductores Léxicos consta de los siguientes elementos:

- un lexicón en forma de FST que define las formas léxicas válidas del lenguaje.
- un conjunto de reglas de dos niveles que asigna la cadena de superficie apropiada de todas las formas léxicas y las categorías morfológicas del lenguaje.

Las reglas son compiladas en traductores y mezcladas con el lexicón utilizando intersección (&) y composición (o) como se muestra en la figura siguiente:

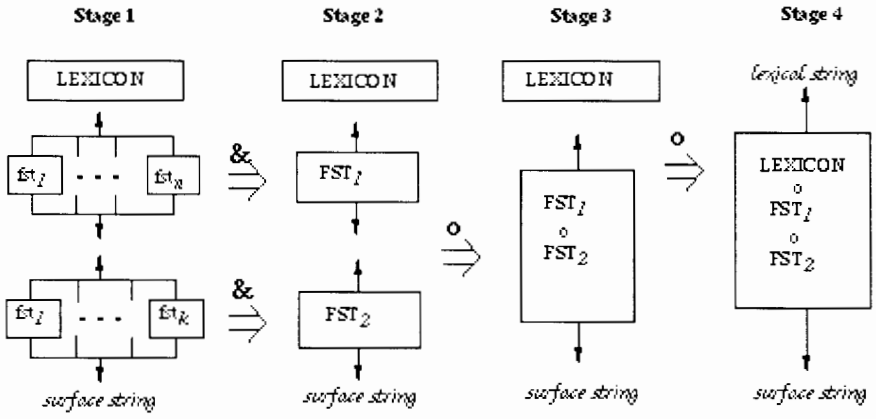


Figura 2: Esquema estándar de construcción de un Traductor Léxico.

En la construcción de estos sistemas los autómatas correspondientes a la intersección de las reglas, alcanza un tamaño considerable, en ocasiones mayor al lexicón, mientras que al componerlas el FST resultante es menor que el que contiene las reglas. Esto sugiere tratar de evitar la representación intermedia de la intersección e implementar una operación que realice al mismo tiempo la intersección de las reglas y su composición con el lexicón. Con este fin Karttunen propuso el algoritmo denominado "intersecting composition" cuya idea básica es construir un FST de salida donde cada estado corresponde a un estado del Lexicón y una configuración de los estados de los Traductores de las reglas. El estado inicial corresponde al estado inicial del Lexicón y su configuración consiste en la secuencia de estados iniciales de los Traductores de las reglas. En cada paso de la construcción se trata de añadir transiciones al estado actual del Traductor de salida mediante el matching de transiciones entre el estado actual del

Lexicón y las transiciones de los estados de todas las reglas incluidos en la configuración actual. Este algoritmo es descrito, aunque de forma bastante superficial, en [Karttunen 1994].

Conclusiones

La morfología de dos niveles es un mecanismo eficiente y suficiente para tratar los diferentes fenómenos flexivos y concatenativos de la morfología del español. No obstante su gran poder expresivo, no deja de ser susceptible a ciertos cambios, sobre todo a partir de las necesidades propias del proyecto que desarrollamos. Estos cambios, que describiremos en el próximo capítulo, estarán dirigidos a la ampliación de la morfotáctica, con la intensión de tratar la dependencia a distancia y la ampliación de las reglas con contextos morfotácticos expresados en función de rasgos morfológicos y clases de palabras.

Capítulo 4.

La construcción del Traductor Léxico

Introducción

El proceso de construcción de un Traductor Léxico necesita de una plataforma para representar el léxico, las reglas y la morfotáctica. En este tema se han desarrollado varios trabajos, [Garrido et. al], [Goñi & González], [Karttunen 1993], [Antworth 1995]. En este capítulo se explican las motivaciones para la creación de una nueva plataforma y se describe el lenguaje y el proceso de compilación utilizado con el objetivo de obtener el Traductor Léxico.

4.1 Consideraciones sobre el Modelo de Dos Niveles.

En varios trabajos se ha hecho referencia al hecho de las limitantes que impone la morfotáctica expresada en términos de clases de continuación, sobre todo, cuando se tratan de expresar las dependencias a distancia entre morfemas. En su lugar se podría utilizar una gramática de palabras basada en los mecanismos de unificación [Antworth 1995], pero su uso impide que la gramática sea convertible en FST y en consecuencia no podríamos obtener el traductor léxico que deseamos.

En [Badía et. al] se habla de la necesidad de que la Gramática de Palabra (GP) sea lo más independiente posible de la representación del léxico y de las reglas. Esto sugiere, en primer término, que la representación de las clases de continuación no se haga directamente en el léxico sino aparte y, tomando en cuanta lo expuesto acerca de sus limitantes como expresión de la morfotáctica, permitir cambiar su forma de representación.

Desde el punto de vista expresivo, la representación de las reglas de dos niveles (RDNs) merece sus comentarios. Según [Badía et. al] "hay que poder restringir la aplicación de las RDNs a ciertas clases de morfemas" y "debería ser posible especificar RDNs de manera que el contexto morfológico sea tenido en cuenta". Intentemos trasladar estas observaciones a la representación del léxico del modelo actual. El término "clases de morfemas" podría ser sustituido por el de "subléxicos". En cuanto al contexto morfológico, en el ámbito de las reglas es posible acceder a los rasgos morfológicos codificados en los FST pero; en primer lugar, los rasgos se codifican al final de cada palabra por lo que para verificar el contexto habría que recorrer toda la palabra para constatar luego si el contexto es válido o no, lo que disminuye la eficiencia del sistema. Adicionalmente las reglas tendrían que tener el orden exacto en que se colocan estos rasgos.

En muchos casos, como se comenta en [Alegría], para poder representar algunos fenómenos y reducir el número de reglas a aplicar se opta por repetir entradas en diferentes subléxicos y esto no queda expresado de forma directa en el modelo. Para resolver este problema en [Goñi & González] se plantea un mecanismo de herencia entre clases de palabras que utiliza los rasgos como mecanismo de control. Este enfoque no sería viable en este modelo por la carencia de una representación clara de los rasgos.

Teniendo en cuenta los criterios expuestos anteriormente se decidió hacer algunas variaciones en cuanto a la representación del léxico y la morfotáctica, ampliando además, las reglas de dos niveles con contextos morfológicos utilizando un formalismo semejante al propuesto en [Badía et. al].

Para la representación del léxico utilizamos clases de palabras en las que se agrupan las palabras con rasgos morfológicos comunes y que asimilan un mismo conjunto de reglas. Para expresar la morfotáctica optamos por una gramática lineal a la derecha y sin ciclos, donde los terminales son las clases de palabras y no hay épsilon-producciones, en su lugar se introduce una clase de palabra denominada "_End" que contiene la marca de fin de palabra.

4.2 El Lenguaje

El lenguaje que proponemos toma como base los trabajos [Garrido et. al], [Goñi & González] y [Karttunen 1993].

Antes de entrar en la descripción del lenguaje primero abordaremos superficialmente el mecanismo de unificación y las estructuras de rasgos que son utilizados en este trabajo para especificar los rasgos morfosintácticos de las clases de palabras y, en el caso de la unificación, como mecanismo de control en la herencia.

4.2.1 Estructuras de rasgos y mecanismos de unificación

Una estructura rasgos (FS) guarda información sobre alguna entidad, usualmente lingüística, como palabra o frase. La información que contiene se expresa en términos de pares atributo-valor.

Una FS es un conjunto de pares atributo-valor, donde el valor puede ser atómico o un FS.

Ej.

```
[ CAT          n              ]
[ AGRMNT       [NUMBER   sg]  ]
[              [PERSON    3]  ]
```

La unificación es una operación binaria definida sobre las FS que falla si en las dos estructuras existe un mismo atributo con valores diferentes, en caso contrario, retorna la unión de las dos estructuras. Ej.

```
[CAT       v]     U     [CAT      n]
[NUMBER   sg]           [PERSON   3]
```

Estas dos estructuras no unifican porque el atributo *CAT* tiene valor *v* en la primera y *n* en la segunda. Ej.

```
[CAT       n]     U     [CAT      n]   =   [CAT       n ]
[NUMBER   sg]           [PERSON   3]       [NUMBER    sg]
                                           [PERSON    3 ]
```

En este caso la unificación si es efectiva.

Este es un formalismo ampliamente empleado en analizadores sintácticos.

4.3.2 Las clases de Palabras

Las clases de palabras son el mecanismo que utilizamos para agrupar un conjunto de palabras que tienen una estructura de rasgo común, que tienen un mismo comportamiento morfotáctico y que aceptan un mismo conjunto de reglas.

Las clases de palabras se definen dentro de los ficheros con extensión ".cls" y su descripción la haremos a partir de un ejemplo:

```
1 class clase1_1 : clase2_1,clase2_2 {
2  features:
3
4   public feat1 = Val1;
5
6   feat2 = {
7     feat2_1 = Val2_1;
8     feat2_2 = val2_2;
9   };
10
11  feat3 = val3;
12
13 words:
14
15   word1;
16   word2 : allomp1;
17   word3;
18   word4 : allomp3;
19
20 rules:
21   regla1;
22
23 rules:
24   regla2 = {
25     []
26     [modo = "Indicativo",tiempo = "pasado"|modo = "subjuntivo",tiempo = "presente"|class4_5]
27   }
28   regla3;
29}
```

Figura 3: Ejemplo de definición de una clase de palabra.

En este ejemplo de definición de una clase de palabra hipotética observaremos en la línea #1 que tiene una sintaxis similar a la empleada en C++. En primer lugar la palabra reservada "class" seguida del nombre de la clase y opcionalmente la lista de clases separadas por coma y encabezadas por el símbolo ":" de las cuales hereda. La herencia se

realiza vía unificación entre la estructura de rasgos propia de la clase y la de cada una de las clases bases, heredándose además todas las palabras y las reglas. En la herencia sólo se consideran las estructuras de rasgos marcadas como "public", como la que se encuentra en la línea #4. A continuación del encabezamiento de la clase pueden definirse las estructuras de rasgos. Cada rasgo puede ser marcado con la palabra "public" indicando que es visible para la herencia. Las palabras que pertenecen a la clase se especifican debajo de la palabra "words" seguida de ":". Cada palabra puede ser marcada con un alomorfo, que se interpretará como la representación de superficie en el FST. Finalmente pueden incluirse bloques de reglas. Cada bloque se encabeza con la palabra "rules" seguida de ":". Y los bloques de reglas se aplicarán en el orden en que se especificaron, aplicando en paralelo sólo las reglas pertenecientes a un mismo bloque. Estas reglas se definen en los ficheros con extensión ".rul" que pueden incluirse en el fichero ".cls" utilizando un mecanismo similar al empleado en C++, por ejemplo "#include <reglas.rul>". Cuando se usan las reglas en una clase en particular se pueden ampliar con los contextos morfológicos izquierdo y derecho. La sintaxis puede ser asimilada a partir del ejemplo expuesto a partir de la línea #24 donde en la regla "regla2" se dice que el contexto izquierdo es vació y que a la derecha debe encontrarse en uno de los casos: [modo = "indicativo", tiempo = "pasado"]. el rasgo *modo* con el valor "indicativo" y el rasgo *tiempo* con el valor "pasado". El operador *or-no exclusivo* expresado con el símbolo "|" permite expresar varios contextos posibles. Dentro de esos contextos las clases de palabras se expresan mediante su identificador. Ej. "class4_5". Para especificar que la regla debe aplicarse siempre que no se encuentre un contexto determinado debe colocarse el símbolo "^" al principio de la descripción del contexto. De no especificarse contexto alguno, la regla se aplica siempre.

4.2.3 Las reglas

En esta etapa de desarrollo no se ha construido el compilador de reglas, por lo que en su lugar permitimos especificar los FSTs en función del alfabeto y la matriz de transición como se muestra en la siguiente figura.

```
1  alphabet alfabeto {
2   surface:
3     a ,b ,c ,d ,e ,f ,g ,h ,i ,j ,k ,l ,m ,n ,ñ ,o ,p ,q ,r ,s ,t ;
4   lexical:
5     a ,b ,c ,d ,e ,f ,g ,h ,i ,j ,k ,l ,m ,n ,ñ ,o ,p ,q ,r ,s ,t ,+ ;
6  }

7  rule r_o_to_a_femenino
8  {
9    alphabet = alfabeto;
10   states = 2;
11   simbols = 3;
12   matriz:
13         o:a  +:0  @:@
14    1:    2    1    1
15    2.    0    1    0
16  }
```

Figura 4: Definición de un FST en función del alfabeto y la matriz de transición.

Las reglas se especificarán en un fichero con extensión ".rul" y la sintaxis, como se pude apreciar en el ejemplo, es bastante simple. Lo primero es la palabra "rule" seguida del identificador que representa el nombre con que se referenciará la regla. Entre llaves se especifican la cantidad de estados y la cantidad de símbolos del alfabeto. Aunque se podrían obviar estos datos es conveniente representarlos para evitar tener que dar dos pasadas sobre el fichero. La matriz de transición se escribe a continuación de la palabra "matriz" y el caracter ":". Los símbolos del alfabeto se escriben por pares y separados por el caracter ":", donde el primero corresponde al nivel de superficie y el segundo al nivel léxico. Para indicar cuales son los estados finales se coloca inmediatamente después el carácter ":" y de no serlo un ".". En el fichero se definen las reglas unas detrás de otras.

El alfabeto que se incluye en la línea 9 se hace necesario para poder limitar los caracteres que se emplean al expandir el símbolo "@" que representa cualquier carácter.

Una descripción detallada de la sintaxis y semántica de las reglas, así como los algoritmos para compilarlas en FST se pueden encontrar en [Antworth 1995].

4.2.4 La gramática de palabra

Ya se mencionó anteriormente que la morfotáctica se representa utilizando una gramática lineal a la derecha y sin ciclos. Estas se definen dentro de un único fichero ".grm" que sirve de base para la compilación del léxico.

Para su descripción, al igual que para las clases de palabras utilizaremos un ejemplo hipotético.

```
1 #include <Class1.cls>
2 #include <Class2.cls>
3 #include <Rules.rul>
4
5 grammar{
6
7  S -> _Prefig PS | _Verbos VS | _Vust SS;
8
9  //la gramática no debe admitir construcciones cíclicas.
10 //del tipo PS ->S;
11
12 PS -> _verbos _Sustantivos VS | _Sust SS;
13
14 VS -> _End | _SufijosVerbales _End;
15
16 SS -> _End | _sufijosNominales _End;
17
18
19 rules:
20  Cegla1;
21  regla2;
22  conjunto1;
23  conjunto3;
24
25 rules:
26  regla1_1;
27  regla1_2;
28}
```

Figura 5: Ejemplo hipotético de definición de una gramática de palabra.

Como se pueden percatar la sintaxis de la definición de las producciones es semejante a la utilizada tradicionalmente. Para definir la gramática se coloca delante la palabra clave

"grammar" y luego entre llaves se escribe la gramática. Las reglas se escriben utilizando una sintaxis y semántica semejantes a la empleada en la declaración de una clase de palabras con la diferencia que en este caso no se permiten incluir contextos morfológicos. En la gramática se pueden incluir conjuntos de reglas entre las clases de palabras, con el objetivo de hacer más evidente el contexto en que se aplican, por ejemplo:

En la línea #7 la primera producción podría tener la forma PS → _Prefig {regla1; regla2;} PS.

4.3 La Compilación

El compilador fue desarrollado utilizando el generador de traductores ANTLR de libre distribución y que permite crear *parsers* descendentes recursivos. Para mayores detalles dirigirse al sitio: http://www.antlr.org/ o directamente al artículo [Terence Parr 2003]. La decisión se tomó teniendo en cuenta las facilidades que brinda este sistema y sobre todo la posibilidad de generar el código para varios lenguajes, entre ellos el C# y que además sus fuentes están disponibles en una librería de clases a las que se les puede hacer modificaciones.

Los pasos para la construcción del FST que se obtiene como resultado de la concatenación y unión de las Traductores correspondientes a las distintas clases de palabras, y la aplicación, en sus contextos, de las reglas que permiten regularizar estas operaciones, se determinan a partir de la especificación de la gramática.

El hecho de que la gramática es lineal a la derecha permite que a partir de ella se pueda obtener un Autómata de Estados Finitos (FSA) y a partir de él, determinar el orden de las operaciones. Convertir la gramática en un FSA nos es una decisión fortuita y tampoco se tomó con el único ánimo de formalizar el proceso. Tener el Autómata permite minimizarlo y en consecuencia eliminar redundancias en la gramática y optimizar el proceso reduciendo el número de operaciones.

4.3.1 El Proceso

Para describir el proceso de construcción del traductor resultante lo haremos a través de un ejemplo sencillo. Supongamos la gramática siguiente:

Gramática:

S -> _Sustantivos SS | _Verbos VS | _PrefijosN PS;

PS -> _End | _Sustantivos SS;

VS -> _End | _SufijosV _End;

SS -> _End | _sufijosN _End;

A partir de ella se obtiene el siguiente autómata:

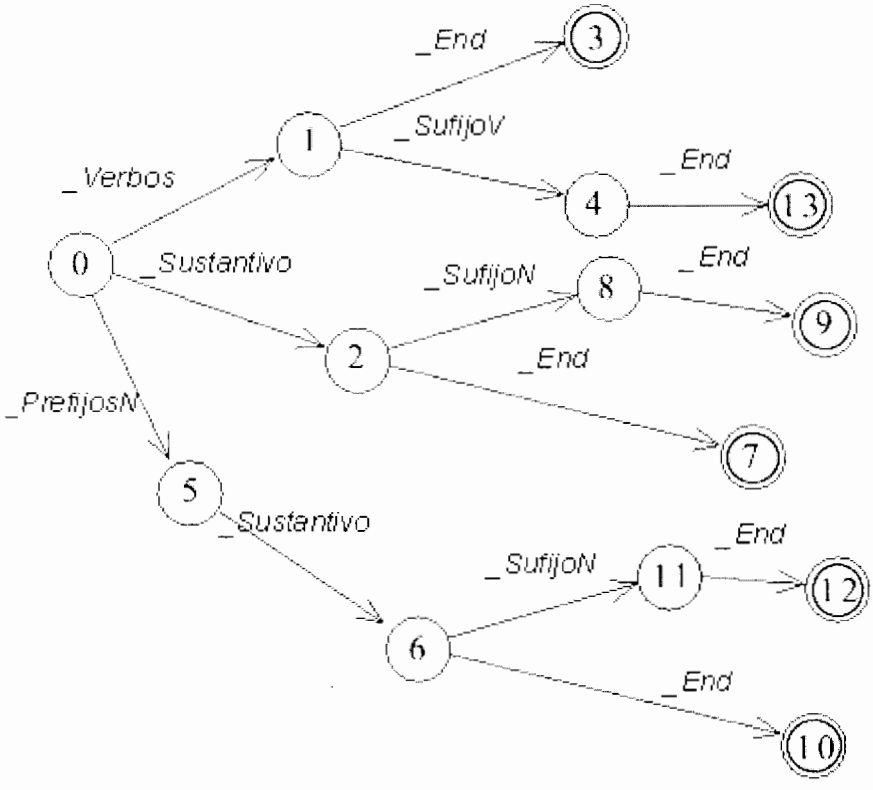


Figura 6. Autómata obtenido a partir de la gramática

Matriz de Transición:

	_Verbos	_End	_Sustantivo	_PrefijosN	_SufijoN	_SufijoV
0.	1		2	5		
1.		3				4
2.		7			8	
3:						
4.		13				
5.			6			
6.		10			11	
7:						
8.		9				
9:						
10:						
11.		12				
12:						
13:						

Tabla 11: Matriz de Transición de un Autómata Finito.

Nótese que este autómata realmente es un árbol que describe todos los caminos posibles en la gramática.

Para disminuir el número de estados y las operaciones de concatenación y unión luego de obtenido el árbol se procede a minimizarlo.

En el algoritmo de minimización se debe incluir la restricción de que para que dos estados puedan compactarse, además de las condiciones habituales, los arcos con un mismo símbolo deben asimilar el mismo conjunto de reglas. Esta restricción adicional, unido a la propiedad de ser un grafo sin ciclos, garantiza que se puedan mantener, luego de minimizado el autómata, los contextos morfológicos originalmente planteados en la gramática.

En este punto es importante hacer notar que el conjunto de reglas de cada arco se obtiene mediante la unión del conjunto de reglas generales y el conjunto de reglas asociadas explícitamente con este en la gramática y en la clase de palabra del símbolo del arco. Para mostrar con más claridad la restricción adicional impuesta a la minimización vamos a

centrarnos en los estados 2 y 6 de la *Figura #6*, los cuales tienen un comportamiento similar, con la diferencia de que en el primero de los casos los sustantivos aparecen al principio de la palabra y en el segundo están precedidos de un prefijo.

Luego de minimizado el autómata tendrá la forma siguiente:

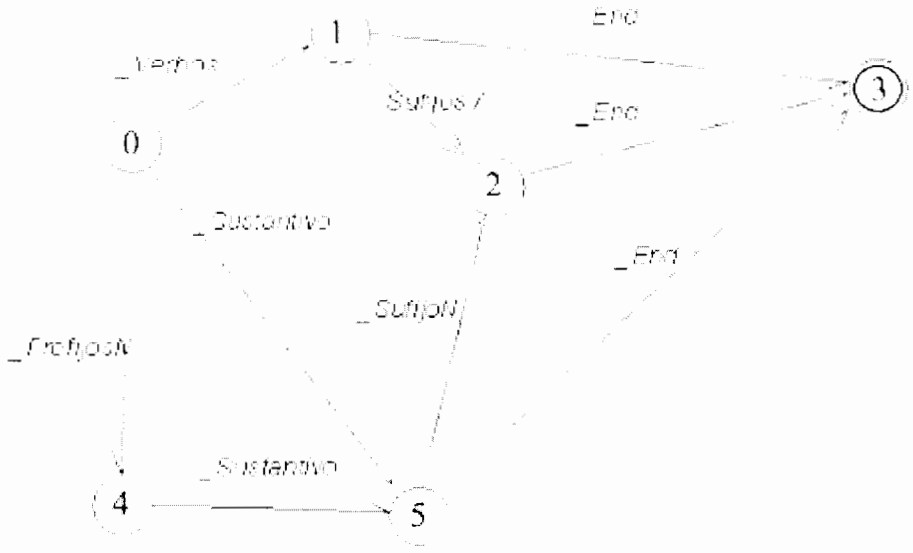


Figura 7. Autómata minimizado

Matriz de Transición:

	_Verbos	_End	_Sustantivo	_PrefijosN	_SufijoN	_SufijoV
0.	1		5	4		
1.		3				2
2.		3				
3:						
4.			5			
5.		3			2	

Tabla 12: Matriz de Transición de un Autómata Finito.

Que, como se puede observar, sufre una reducción considerable en su número de estados.

Luego, para obtener el orden de las operaciones bastaría con resolver el sistema:

$$FST0 = \bigcup_{i=0}(\_Verbos + FST1_i) \cup \bigcup_{i=0}(\_Sus\tan tivos + FST5_i) \cup \bigcup_{i=0}(\_\Pr efijosN + FST4_i)$$

$$FST1 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right), \left(\bigcup_{i=0}(\_SufijosV + FST2_i)\right)\right\}$$

$$FST2 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right)\right\}$$

$$FST3 = \{(\phi)\}$$

$$FST4 = \left\{\left(\bigcup_{i=0}(\_Sus\tan tivos + FST5_i)\right)\right\}$$

$$FST5 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right), \left(\bigcup_{i=0}(\_SufijosN + FST2_i)\right)\right\}$$

Ahora bien, esta minimización se realizó bajo el supuesto de que los arcos iguales que parten de estos dos estados asimilaran el mismo conjunto de reglas. Supóngase que para el arco marcado con el símbolo "*_SufijoN*" que parte del estado 6 en la *Figura #6*, cuando el sustantivo está precedido de un prefijo, se quisiera aplicar alguna regla adicional que tenga en consideración el hecho de que los sustantivos están precedidos de un prefijo. En este caso los estados no se podrían reducir y el autómata minimizado tendría la forma que se muestra en la Figura # 8.

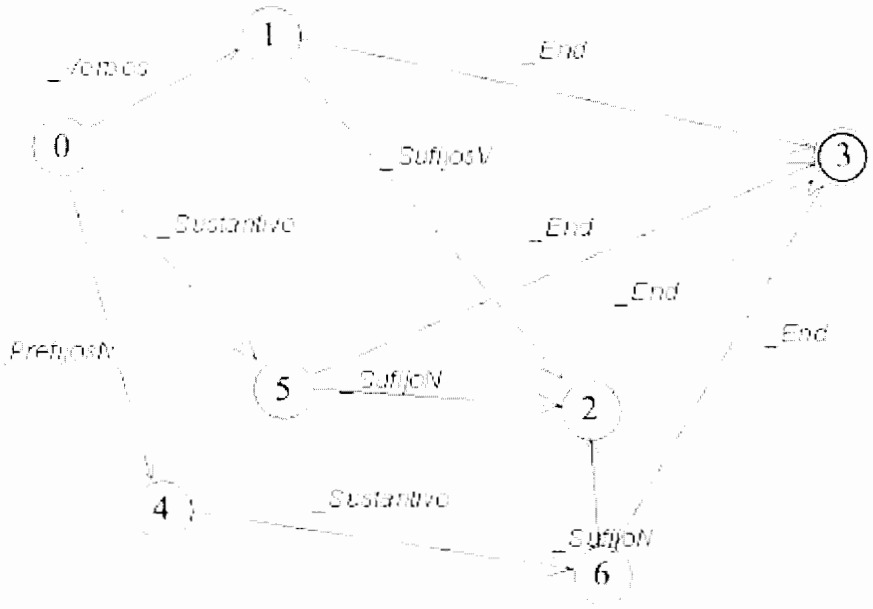


Figura 8. Autómata minimizado teniendo en cuenta las restricciones de los estados

Matriz de Transición:

	_Verbos	_End	_Sustantivo	_PrefijosN	_SufijoN	_SufijoV
0.	1		5	4		
1.		3				2
2.		3				
3:						
4.			6			
5.		3			2	
6.		3			2	

Tabla 13: Matriz de Transición de un Autómata Finito.

Y el sistema a resolver sería:

$$FST0 = \bigcup_{i=0}(\_Verbos + FST1_i) \cup \bigcup_{i=0}(\_Sus\tan tivos + FST5_i) \cup \bigcup_{i=0}(\_\Pr efijosN + FST4_i)$$

$$FST1 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right), \left(\bigcup_{i=0}(\_SufijosV + FST2_i)\right)\right\}$$

$$FST2 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right)\right\}$$

$$FST3 = \{(\phi)\}$$

$$FST4 = \left\{\left(\bigcup_{i=0}(\_Sus\tan tivos + FST6_i)\right)\right\}$$

$$FST5 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right), \left(\bigcup_{i=0}(\_SufijosN + FST2_i)\right)\right\}$$

$$FST6 = \left\{\left(\bigcup_{i=0}(\_End + FST3_i)\right), \left(\bigcup_{i=0}(\_SufijosN + FST2_i)\right)\right\}$$

Conclusiones

En este capítulo se ha descrito el lenguaje para representar el léxico de nuestra lengua tomando como base la morfología de dos niveles, además se mostraron detalles del proceso de compilación de este léxico en un Traductor utilizando una gramática lineal a la derecha y sin ciclos.

Conclusiones

Los *Sistemas de Consulta en Lenguaje Natural a Ficheros de Preguntas Frecuentes* son una opción eficiente para facilitar el acceso a lo que puede considerarse, en todos los aspectos, como una base de conocimiento importante. ¿Ventajas? Muchas. Tan solo considerar una Base de Datos en donde se introduce conocimiento sobre temas múltiples en un formato tan simple como el que utilizamos para comunicarnos es un motivo suficiente para su valoración. Los resultados sólo se han obtenido a nivel de prototipos, pero sin dudas son una opción importante a considerar.

Para lograr un producto robusto se necesita de varias herramientas y el mérito estará en lograr una medida justa entre el poder de análisis y el tiempo de respuesta. Para extender la profundidad del análisis es importante integrar herramientas con un nivel de eficiencia significativo. Puntos críticos son el análisis morfológico y el acceso a los ficheros de índices con toda la información posible para describir la pregunta. Aunque de forma general todas las herramientas: desambiguador morfosintáctico, analizador sintáctico y el módulo para el tratamiento de sinónimos e hiperónimos tienen gran responsabilidad en el proceso.

Como consideraciones importantes a exponer al concluir este trabajo están las siguientes:

- En este trabajo se realizó un estudio profundo de la morfología de nuestra lengua materna, a partir de conceptos lingüísticos y formalismos computacionales. Los resultados se centran en un lenguaje para la representación del léxico mediante *clases de palabras* con características similares desde el punto de vista morfotáctico y que comparten los mismos rasgos morfosintácticos.
- Otro resultado obtenido es la representación de la morfotáctica utilizando como base una gramática lineal a la derecha y sin ciclos que permite el tratamiento de la dependencia a distancia de los morfemas y mantiene la propiedad de ser convertible en un Traductor de Estados Finitos, utilizando el algoritmo de

minimización para optimizar la cantidad de operaciones a realizar. La expresividad del lenguaje se ve premiada con la ampliación de las reglas de dos niveles con contextos morfológicos expresados en términos de clases de palabras y rasgos morfosintácticos y, además, con la posibilidad de incluir directamente conjuntos de reglas en puntos específicos de la gramática.

- La construcción de un módulo completo para evaluar y validar los cambios propuestos en el proceso de construcción del Traductor Léxico no se ha concluido por lo que no se puede arribar a valoraciones concluyentes, no obstante el estado de su desarrollo muestra su viabilidad.

En términos generales este ha sido un trabajo en el que, principalmente, se lanzan interrogantes y criterios importantes a tener en cuenta para la construcción del sistema final.

Recomendaciones

Después de concluir este trabajo se recomiendan acometer las siguientes acciones:

- Profundizar en el estudio de los desambiguadores morfosintácticos que concluya con la implementación de un módulo con esta funcionalidad.
- Profundizar en el estudio de los analizadores sintácticos superficiales que concluya con la implementación de un módulo con esta funcionalidad.
- Reclamar del concurso de un especialista en lingüística para, en conjunto, determinar el mínimo de información necesaria para indexar las preguntas en la Base de Conocimientos.
- Concluir la implementación del procesador morfológico.
- Evaluar la posibilidad y utilidad de la inclusión de la relación de meronimia en la ontología.
- Definir una estructura para la representación y gestión de los sinónimos e hiperónimos en la ontología.
- Definir una medida cuantitativa para el grado de cercanía entre dos preguntas.
- Construir el sistema de consulta en lenguaje natural a ficheros de preguntas frecuentes a partir de las herramientas desarrolladas.

Referencias Bibliográficas

- [Alegría] Alegría Iñaqui. *Morfología de Estados Finitos.* Informatika Facultatea (UPV / EHU).
- [Antworth 1995] Antworth, Evan L. *User's Guide to PC-KIMMO Version 2.* Summer Institute of Linguistics. (http://www.sil.org/pckimmo/v2/doc/guide.html)
- [Badia et. al.] Badia T., Egea A. Tuells A. *SEGMORF: un formalismo para analizadores morfológicos de dos niveles,* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.
- [Burger et. al. 2001] Burger Jhon. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST. 2001
- [Burke et. al. 1997] Burke R., Hammon K., Kulyukin V., Lytinen S., Tomuro N., Scoenberg S. *Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System.* 1997.
- [Carulla Oosterho] Carulla M. Oosterho A., *El tratamiento de la morfología flexiva del castellano mediante reglas de dos niveles en una gramática de unificación.* Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona.
- [Garrido et. al] Garrido Alicia, Iturraspe Amaia, Montserrat Sandra, Pastor Hermínia, Forcada Mikel L. *A compiler for morphological analysers and generators based on finite-state transducers.* Departament de Llenguatges i Sistemes Informátics, Universitat d'Alacant, Spain.
- [Goñi & González] José M. Goñi & José C. González. *A framework for lexical representation.* E. T. S. I. Telecomunicación, Universidad Politécnica de Madrid.
- [Karttunen 1993] Karttunen Lauri. *Finite-State Lexicon Compile.* Xerox Palo Alto Research Center. Technical Report. ISTL-NLTT-1993-04-02. April 1993.

- [Karttunen 1994] Karttunen Lauri. *Constructing Lexical Transducers.* Xerox Research Centre Europe, Grenoble. In The Proceedings of the 15th International Conference on Computational Linguistics. Coling 94. Kyoto, Japan 1994. (http://www2.parc.com/istl/members/karttune/publications/coling-94/coling94.html)

- [Light et. al. 2001] Light Marc, Mann Gideon, Riloff Ellen & Breck Eric. *Analyses for Elucidating Current Question Answering Technology.* Natural Language Engineering. Cambridge University Press. 2001

- [Mills 1999] Mills, Jon., *A logical approach to the lemmatization of computational lexica.* Actas-1 del VI Simposio Internacional de Comunicación Social, Santiago de Cuba, Cuba, 1999.

- [Mollá et. al. 1998] Mollá Aliod Diedo, Berri Jawad, Hess Michael. *A Real World Implementation of Answer Extraction.* Proceeding of the 9th International Workshop on Database and Expert Systems Applications Workshop "Natural Language and Information Systems" (NLIS?98). Vienna, Australia. August 26-28, 1998

- [Moreno] Moreno, Antonio., *GRAMPAL: A Morphological Processor for Spanish implemented in Prolog,* Dep. de Lingüística Universidad Autónoma de Madrid, España.

- [Potier 1970] Bernard Pottier, Ligüística Moderna y Filología Hispánica. Editorial Gredos, Madrid, 1970.

- [Roche E. & Schabes Y. 1996] Roche Emmanuel & Schabes Yves. *Introduction to Finite-State Devices in Natural Language Processing.* Mitsubishi Electronic Research Laboratories. 1996.

- [Rodríguez & Carretero 1994] Rodríguez Santiago, Carretero Jesus., *A Formal Approach to Spanish Morphology the COES Tools.* Facultad de Informática. Univ. Politécnica de Madrid. 1994

- [Rodríguez E.] Rodríguez Vázquez de A. Emilio. Indexación con aproximación lingüística. Un breve repaso a los "intentos de mejora" en el proceso de indexación. Universidad de Salamanca. Grupo REINA. España.
- [Sneiders] Sneiders Eriks: *Automated FAQ Answering: Continued Experience with Shallow Language Understanding.* Department of Computer and System Science, Stockholm University / Royal Institute of Technology, Sweden.
- [Tou et. al. 2001] Ng Tou Hwee, Pheng Kwan Lai Jennifer & Xia Yiyuan. *Question Answering Using a Large Text Database: A Machine Learning Approach.* Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP2001). Pittsburgh, PA. June 3-4, 2001
- [Vicedo & Ferrández 2000] Vicedo José Luis & Ferrández Antonio. *A semantic approach to Question Answering Systems.* In TREC-9. Nineth Text Retrieval Conference, 2000
- [Vilares et. al. 2001] Vilares J., Cabrero D., Alonso M., *Generación automática de familias morfológicas mediante morfología derivativa productiva.* Revista: Procesamiento del Lenguaje Natural. España, vol. 27, 2001

Bibliografía

1. Alegría Iñaqui. *Morfología de Estados Finitos.* Informatika Facultatea (UPV / EHU).
2. Ampuero Alberto Juan, *Lengua y Literatura Castellana.* (http://mimosa.pntic.mec.es/~ajuan3/otraswww/r_teor_4.htm)
3. Antworth, Evan L. *User's Guide to PC-KIMMO Version 2.* Summer Institute of Linguistics. (http://www.sil.org/pckimmo/v2/doc/guide.html)
4. Badia T., Egea A. Tuells A. *SEGMORF: un formalismo para analizadores morfológicos de dos niveles,* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona,
5. Beniers Jacobs, Elizabet., *La noción de productividad vista en relación con la derivación española.* Univ. Nacional Autónoma de México. Imprenta Universitaria 1985.
6. Burke R., Hammon K., Kulyukin V., Lytinen S., Tomuro N., Scoenberg S. *Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System.* 1997.
7. Burger Jhon. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST. 2001
8. Light Marc, Mann Gideon, Riloff Ellen & Breck Eric. *Analyses for Elucidating Current Question Answering Technology.* Natural Language Engineering. Cambridge University Press. 2001

9. Carulla M. Oosterho A., *El tratamiento de la morfología flexiva del castellano mediante reglas de dos niveles en una gramática de unificación.* Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona.

10. Cutting, D., J. Kupiec, J. Pedersen and P. Sibun. *A Practical Part-of-Speech Tagger*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, p. 133-140, 1992

11. Garrido Alicia, Iturraspe Amaia, Montserrat Sandra, Pastor Hermínia, Forcada Mikel L. *A compiler for morphological analysers and generators based on finite-state transducers.* Departament de Llenguatges i Sistemes Informátics, Universitat d'Alacant, Spain.

12. José M. Goñi & José C. González. *A framework for lexical representation.* E. T. S. I. Telecomunicación, Universidad Politécnica de Madrid.

13. Karttunen Lauri. *Finite-State Lexicon Compile.* Xerox Palo Alto Research Center. Technical Report. ISTL-NLTT-1993-04-02. April 1993.

14. Karttunen Lauri. *Constructing Lexical Transducers.* Xerox Research Centre Europe, Grenoble. In The Proceedings of the 15th International Conference on Computational Linguistics. Coling 94. Kyoto, Japan 1994. (http://www2.parc.com/istl/members/karttune/publications/coling-94/coling94.html)

15. Mathias Creutz. *Morphology and Finite-State Transducers. Chapter 3, Jurafsky & Martin.* 31 Octubre 2001. (http://www.cis.hut.fi/Opinnot/T-61.184/s01/mathias.ppt)

16. Mills, Jon., *A logical approach to the lemmatization of computational lexica.* Actas-1 del VI Simposio Internacional de Comunicación Social, Santiago de Cuba, Cuba, 1999.

17. Mollá Aliod Diedo, Berri Jawad, Hess Michael. *A Real World Implementation of Answer Extraction.* Proceeding of the 9th International Workshop on Database and Expert Systems Applications Workshop "Natural Language and Information Systems" (NLIS?98). Vienna, Australia. August 26-28, 1998

18. Moreno, Antonio., *GRAMPAL: A Morphological Processor for Spanish implemented in Prolog,* Dep. de Lingüística Universidad Autónoma de Madrid, España.

19. Ochoa Luis, Gimenez Mico José A. *SPANISH 301. Gramática y Composición. Fichas Gramaticales.* Otoño 2001. (http://artsandscience.concordia.ca/cmll/spanish/ochoa/301_Fichas_gramaticales.htm)

20. Pérez González, Graciela., *Los prefijos en el DRAE y en algunos diccionarios cubanos.* Editorial Academia, La Habana, 1988

21. Bernard Pottier, Ligüística Moderna y Filología Hispánica. Editorial Gredos, Madrid, 1970.

22. Roche Emmanuel & Schabes Yves. *Introduction to Finite-State Devices in Natural Language Processing.* Mitsubishi Electronic Research Laboratories. 1996.

23. Rodríguez Vázquez de A. Emilio. Indexación con aproximación lingüística. Un breve repaso a los "intentos de mejora" en el proceso de indexación. Universidad de Salamanca. Grupo REINA. España.

24. Rodríguez Santiago, Carretero Jesus., *A Formal Approach to Spanish Morphology the COES Tools.* Facultad de Informática. Univ. Politécnica de Madrid. 1994

25. Sneiders Eriks: *Automated FAQ Answering: Continued Experience with Shallow Language Understanding.* Department of Computer and System Science, Stockholm University / Royal Institute of Technology, Sweden.

26. Terence Parr. *ANTLR Referente Manual*

27. Civit Torruella, Montserrat & Martí Antonin, Ma. Antonia. *Análisis Morfosintáctico para la extracción de información.* CLiC Centre de Llenguatge i Computació.

28. Ng Tou Hwee, Pheng Kwan Lai Jennifer & Xia Yiyuan. *Question Answering Using a Large Text Database: A Machine Learning Approach.* Proceedings of the

2001 Conference on Empirical Methods in Natural Language Processing (EMNLP2001). Pittsburgh, PA. June 3-4, 2001

29. Cárdenas Urrutia, Hernán. Alvarez Alvarez, M. *Esquema de Morfosintaxis Histórica del Español.* Publicaciones de la Univ. De Deusto, Bilbao, 1998.

30. Velásquez Francisco, Ggelbukh Alexander, Sidorov Grigori. *AGME: Un sistema de Análisis y Generación de la Morfología del Español.* Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN). México.

31. Vicedo José Luis & Ferrández Antonio. *A semantic approach to Question Answering Systems.* In TREC-9. Nineth Text Retrieval Conference, 2000

32. Vilares J., Cabrero D., Alonso M., *Generación automática de familias morfológicas mediante morfología derivativa productiva.* Revista: Procesamiento del Lenguaje Natural. España, vol. 27, 2001