Un algoritmo para la extracción de rasgos morfológicos a partir de descriptores

Juan José Prada
Instituto de Computación - Facultad de Ingeniería
Universidad de la República - URUGUAY
Agosto 1996
e-mail : prada@fing.edu.uy
jprada@nirf.imm.gub.uy

En este trabajo se propone un algoritmo para la alimentación de un diccionario español a partir de un conjunto de términos para la indización de documentos (descriptores) organizados en un tesauro. Se establecen e implementan heurísticas basadas en la estructura sintáctica de los descriptores y en la forma de las palabras que permiten deducir los siguientes rasgos morfológicos: categoría gramatical (sustantivo, adjetivo), número, género y formas flexionadas.

Se utiliza como fuente de datos un tesauro y un diccionario mínimo de palabras "vacías" (conjunto base) tales como preposiciones, artículos, etc.

Términos clave: morfología, descriptor, tesauro, procesamiento de lenguaje natural

ÍNDICE

0 INTRODUCCIÓN

En este trabajo se propone un algoritmo para la alimentación de un diccionario español a partir de un conjunto de términos para la indización de documentos (descriptores) organizados en un tesauro. Se establecen e implementan heurísticas basadas en la estructura sintáctica de los descriptores (grupos nominales) y en la forma de las palabras que permiten deducir los siguientes rasgos morfológicos:

- categoría gramatical (sustantivo, adjetivo)
- género
- número

Se utiliza como fuente de datos un tesauro y un diccionario mínimo de palabras "vacías" (conjunto base) tales como preposiciones, artículos, etc.

El documento se estructura del siguiente modo :
En una primera parte, que abarca los primeros cuatro capítulos se establece el marco general del trabajo. En el capítulo 1, se exponen algunas de las motivaciones por las cuales se realizó el trabajo, presentando en el capítulo 2 algunos antecedentes de trabajos con puntos de contacto con el nuestro. En el capítulo 3 se hace un breve estudio de la morfología del español, haciendo énfasis en la estructura de los grupos o sintagmas nominales, reduciendo luego este análisis al tratamiento de las estructuras que presentan los términos del tesauro. En el capítulo 4 se brindan algunos conceptos generales de informática documental.

Una segunda parte, que comprende los capítulos 5, 6 y 7 contiene la propuesta del trabajo, su implementación y las conclusiones. En los capítulos 5 y 6 se plantea el conjunto de heurísticas y se propone una solución informática para obtener los rasgos mencionados para cada palabra reconocida, analizándose además los resultados obtenidos. Dichos resultados se obtuvieron a partir de un conjunto de datos de prueba (1100 entradas que corresponden a un análisis de 1200 palabras distintas aproximadamente), y el éxito fue del orden del 95%, lo cual es considerado altamente satisfactorio. En el capítulo 7, se proponen una serie de ampliaciones a este trabajo a los efectos de mejorar la precisión de los resultados obtenidos y de posibilitar la integración a un sistema de información automatizada.

Por último se adjunta la bibliografía manejada y un apéndice conteniendo los fuentes de la implementación[1].

[1] En este reporte, el apéndice está formado por a1 y a2, dejando de lado los fuentes de la implementación

1 MOTIVACIÓN

Uno de los mayores problemas en las aplicaciones asociadas al procesamiento del lenguaje natural, es la adquisición de conocimiento sobre un dominio específico. Este conocimiento es fundamental debido a la tendencia actual de permitir que el usuario formule sus requerimientos en un lenguaje libre. Los sistemas informáticos que interpretan estos requerimientos, deben disponer de extensas Bases de Conocimiento sobre el dominio específico. En este sentido, se han desarrollado herramientas para la alimentación automática de estas bases a partir de otros datos presentes en la aplicación [COC91].

Por ejemplo, en la construcción de interfases en lenguaje natural a Bases de Datos Relacionales (BDR), se construyen alimentadores que utilizan la descripción de la BDR, bajo la hipótesis de que se utilizan nombres significativos en tablas y atributos.

El uso de tesauros es una excelente fuente de información y asistencia para el usuario que plantea una consulta en lenguaje libre, puesto que los tesauros contienen un conjunto de conceptos que ayudan tanto a indizar como a recuperar los datos (o documentos) al establecerse relaciones entre los términos similares allí presentes.

Para una aplicación concreta, en la que se trabaja en un dominio específico y muy técnico, suelen existir en los tesauros, palabras que no se encuentran en un diccionario tradicional. En este sentido, el tesauro constituye una excelente fuente de alimentación para un diccionario de la aplicación. Por otra parte, es de destacar que la alimentación manual de un diccionario es una tarea engorrosa.

En este artículo se propone una solución computacional a dicho problema a través de una serie de heurísticas que determinen para cada palabra la información de género, número y categoría gramatical.

2 ANTECEDENTES

Existen diversas publicaciones vinculadas a la extracción de rasgos de las palabras, apoyándose en diccionarios de distinto tipo y tamaño.

En [OGO93] se construye un parser bottom-up con la ayuda de un gran lexicón y realizando predicciones para las categorías de palabras desconocidas. Nuestra propuesta, por el contrario, se basa en un conjunto mínimo de palabras (preposiciones, conjunciones, determinantes, etc.) y se propone inferir la categoría gramatical para el resto de las palabras.

Por otro lado, es de destacar que la entrada de datos para nuestro análisis la constituyen grupos nominales y no oraciones, como en la propuesta de [VER94], donde se construye un parser de complejidad lineal que distingue entre palabras y oraciones, estableciendo estructuras intermedias denominadas "secuencias" - nominales o verbales - donde la palabra central es un nombre o un verbo según el caso. Si bien éste parser fue realizado para el francés con versiones para el español e inglés, existe en ese análisis de estructuras un punto importante de contacto con nuestra propuesta. La diferencia es que el trabajo de [VER94] se apoya en un diccionario de raíces verbales, con el que nuestro trabajo no cuenta.

Además de la categoría gramatical, nos interesa obtener los rasgos de género y número para cada palabra. Al respecto, podemos citar el trabajo de [SAN95] que propone un algoritmo de singularización, apoyándose en un diccionario para resolver algunos problemas que pudieran aparecer durante el proceso. Es de destacar que el diccionario empleado en este trabajo excluye los verbos, las preposiciones adverbios, etc. y además toma en cuenta la acentuación de las palabras. Nuestro trabajo no contempla la acentuación ya que no dispone de ésta información debido a que la fuente de datos utilizada - Tesauro Spines - está escrita con letras mayúsculas.

3 ALGUNOS CONCEPTOS DE MORFOLOGÍA

3.1 Introducción

Si bien no se pretende en este documento hacer un análisis profundo de la morfología del español, se considera pertinente comentar algunos conceptos que son tratados y que determinan el marco de trabajo.

La morfología (del griego *morfe* : forma), estudia la forma de las palabras, a sus accidentes (flexión, composición y derivación) y a su clasificación en <<partes de la oración>>. En un estudio morfológico se obtienen las palabras de un contexto, se las clasifica según su función gramatical, se estudia la forma que pueden adquirir para representar las distintas categorías.

3.2 La palabra

"*La palabra es el centro del universo lexical*" [CAL83].

La *palabra*, es el elemento lingüístico compuesto por una secuencia de caracteres generalmente individualizadas por un carácter especial : el separador. En la escritura tradicional, este separador se corresponde con el espacio en blanco, aunque puede ser coma (,), punto y coma (;), guión (-), etc.

Es importante destacar, que una misma secuencia de caracteres, puede tener distintos significados según su contexto y esto sin lugar a dudas puede llegar a ocasionar problemas cuando se la analiza aisladamente.

El morfema

El ***morfema*** es la mínima unidad con significado, es decir una secuencia de elementos asociados a un significado que no pueden ser divididos en unidades significativas menores.

"*El morfema puede ser asociado muchas veces con una palabra : sol , mar, siempre. Hablamos entonces de palabras radicales. En otros casos el morfema es parte de la palabra, como en sol-ar (adjetivo), carcel-ero, mar-es ...*" [RAE86]

Por su parte, los morfemas se pueden clasificar en:

1	2	3	4
Prefijos	**Base**	**Sufijos**	**Desinencias**
in	cont	able	s
in	habil	idad	0
co	oper(a)	ción	0
	cant(a)		ba
	cant		é

- las bases aportan el significado léxico
- las desinencias tienen un significado exclusivamente gramatical, indican género, número, tiempo, persona, etc. de los acompañantes

Categorías de palabras

Las palabras se pueden agrupar en clases o categorías :

- **sustantivo** : refieren a objetos, personas : casa, Juan, universidad, etc.

- **adjetivo** : calificativo :dan las cualidades esenciales o circunstanciales de los seres : bueno, malo, etc.

determinativo : acompañan al nombre (al que califican) aportando datos para su correcta localización : estos, varios, etc.

- **artículo** : el , las, etc.

- **pronombre** :palabra con significado referencial, que en cada caso significa cosas distintas pero que suele representar a un nombre : tu, nosotros, etc.

- **verbo** : palabra que expresa acciones : comer, amar, etc.

- **adverbio** : palabra matizadora de la acción verbal, del adjetivo, de otro adverbio, o de toda una oración : mucho, muy, rápidamente, etc.

- **preposición** y **conjunción** : elementos de enlace entre otras clases de palabras : de, con, para, y, etc.

Rasgos

Características que presentan cada una de las palabras:
- género : masculino, femenino, invariante.
- número : singular , plural.
- modo : indicativo, imperativo, etc.
- persona : primera, segunda, tercera.
- tiempo : presente, futuro, etc.

Más adelante, en este documento se analiza como se combinan algunos de estos rasgos según la categoría de la palabra.

3.3 Unidades sintácticas

En este punto se introduce brevemente algunos conceptos básicos de la Sintaxis (del griego *sintasso* : establecer un orden), ya que en el trabajo se estudia una de las estructuras sintácticas formadas a partir del agrupamiento de palabras.

La separación entre morfología y sintaxis es arbitraria [RAE86], y solo se basa en una conveniencia al estudiar la lengua desde distintos puntos de vista. Esto hace que en ocasiones se hable de ***morfosintaxis,*** es decir que considera a los hechos y fenómenos del lenguaje en su dimensión formal y funcional a la vez.

Las palabras, mediante ciertos criterios - semántico : por ejemplo entre el sustantivo y el verbo ; posicional : por ejemplo entre el adjetivo y la preposición, etc.-, llegan a conjuntarse y forman nuevas unidades léxicas más complejas : ***sintagmas***. Estos se organizan en torno a una palabra (denominada *núcleo del sintagma*) y dependiendo de la categoría léxica de esta, existen distintos tipos de sintagmas (oraciones, nominales, verbales, etc.).
Al conjunto de estos criterios, se les denomina ***Gramáticas***, y se puede decir que la *oración* es la unidad superior percibida con más claridad por el hablante no solo como una unidad estructural, sino también como una unidad con significado propio.

"*La oración es la unidad más pequeña de sentido completo en sí misma en que se divide el habla real* " [RAE86].

Cabe decir que esta definición es la proporcionada por [RAE86] aunque es discutida por algunos gramáticos, quienes afirman que existen otras unidades más pequeñas que la oración que también tienen sentido.

De todas maneras, al definirla como menor unidad, no debe pensarse en su extensión; lo importante es que cada una tenga *sentido completo en si misma* y exprese una pregunta, mandato, deseo, etc.
Cuando la morfología clasifica las palabras como partes de la oración, se basa en conceptos sintácticos; cuando la sintaxis establece las reglas de concordancia, se basa en el análisis morfológico hecho de la forma de cada palabra.

Se puede decir entonces, que el morfema y la oración son las unidades extremas en las gramáticas. Si bien la oración se constituye de una secuencia de palabras, esta secuencia se puede dividir en unidades intermedias con estructura propia.
Por ejemplo, en la oración

Las interfases en lenguaje natural son de gran ayuda

claramente, el sujeto lo constituye la expresión "*Las interfases en lenguaje natural* ", la que se denomina ***frase*** *(**grupo/sintagma**)* ***nominal***, que es la unidad léxica que tiene al sustantivo "*interfases*" como núcleo y que a su vez posee una estructura interna. Es decir, que la función sintáctica de sujeto no la realiza necesariamente una palabra, sino una unidad lingüística que posee un núcleo (*interfases*) y complementos de ese núcleo (*en lenguaje natural*) que aportan cierta información específica.
Aparecen entonces otro tipo de unidades intermedias que se denominan ***sintagmas***.

Los *sintagmas* son unidades de construcción que casi todas las escuelas gramaticales consideran fundamentales en el estudio de la sintaxis. En la lingüística moderna aparecen términos tales como *sintagma nominal, sintagma verbal, sintagma adjetival,* etc. Bosque, en [BOS91] cita a Bello, quien propone que de la misma manera que se distinguen las palabras en clases, hacer lo propio con los sintagmas. De esta forma, un sustantivo con las modificaciones que lo especifican o explican (*interfases en lenguaje natural*) forma un ***sintagma nominal*** (**SN** o frase sustantiva) a la cual es aplicable todo lo que dice el sustantivo.
De la misma forma, un verbo con sus respectivos modificadores conforma un ***sintagma verbal*** (**SV** o frase verbal); un adjetivo con los suyos forma un ***sintagma adjetival*** (**SA** o frase adjetival); etc.

No se va a continuar profundizando en el estudio de cada una de estas unidades.
En el siguiente punto se trata un poco más en detalle los sintagmas nominales, puesto que este trabajo se centra en la estructura de éstos debido a que las entradas del tesauro consisten de este tipo de unidades.

3.4 El sintagma nominal

Como ya se expresó anteriormente, un SN es la unidad léxica formada por un nombre (sustantivo) más un conjunto de palabras que lo "acompañan" o complementan.
Su estructura es:

Determinante - Núcleo - Complementos

que es su forma plena, aunque también se puede presentar en alguna de sus formas reducidas,

Determinante - Núcleo
Núcleo
Núcleo - Complementos

Si bien en general se lo identifica con el *sujeto* de una oración, un SN puede cumplir en éstas cualquier función que pudiera desempeñar el sustantivo, por ejemplo :

Juan le dio <u>flores</u> a su mamá	- Complemento directo
Juan le dio flores a <u>su mamá</u>	- Complemento indirecto : prep + SN
Montevideo, <u>capital del Uruguay</u>	- Aposición
Voy al recital con <u>los amigos de Mónica</u>	- Complemento circunstancial : prep + SN
El fax fue enviado por <u>la secretaria</u>	- Complemento agente : prep + SN

A continuación, se analizan los componentes de un SN.

Determinantes

Son las palabras que acompañan al sustantivo en el SN y concuerdan con él en los rasgos género y número. Están en esta clase [MAD89]:

- Artículos
  - determinados : **la, el, lo, las, los**
  - contractos, formados a partir de la contracción producto de una preposición y un artículo determinado : **al, del**
  - indeterminados : **un, una, unos, unas**

- Adjetivos determinativos
  - posesivos : indican posesión y pueden ir antepuestos o pospuestos al sustantivo como son los casos de : **mi, mío, tu, suyo**, etc.
  - demostrativos : indican proximidad, lejanía desde el punto de vista del hablante, como es el caso de : **este, esa, aquellos**, etc.
  - indefinidos : presentan al sustantivo sin concretarlo : **algún, demasiado**, **mucho**, etc.
  - relativo : presentan un valor posesivo : **cuyo, cuyas**, etc.
  - interrogativos / exclamativos : acompañan al sustantivo en frases interrogativas o exclamativas : **qué, cuál, cuánto**, etc.

Núcleos

El núcleo del SN, es por definición el ***sustantivo***, pero pueden funcionar como núcleos también:

un pronombre,	*La que me dio María*
un infinitivo,	*Dormir , es mejor que estar despierto*
una palabra sustantivada,	*La comida de la cantina no es muy sabrosa*

De todas maneras, estos tres casos, no serán analizados en este documento ya que el estudio se basó solo en el análisis de los nombres sustantivos, por ser la categoría que aparece en los SN del tesauro de prueba.

- Nombre Sustantivo.-

El nombre sustantivo, llamado también únicamente ***sustantivo***, es la parte del SN que designa seres, personas o cosas con existencia independiente, ya sea en la realidad (máquina, casa), en la abstracción (crisis, virtud) o en la personificación (Juan).
Recién en la 12ª edición de la gramática de la RAE en 1870, se acepta al nombre sustantivo como una categoría distinta del ***nombre adjetivo*** (que se trata más adelante). Si bien al día de hoy el sustantivo y el adjetivo aparecen como categorías independientes, poseen muchas características comunes; por ejemplo tienen los mismos morfemas de número y en muchos casos también coinciden en los sufijos.

En varias ocasiones, es la construcción sintáctica la que decide la categoría.
Esta observación es crucial puesto que va a constituir una regla importante en el desarrollo de este trabajo.
No obstante ello, es importante destacar también que esto no siempre ocurre, o mejor dicho, que existen palabras que pueden cumplir las dos funciones gramaticales según el contexto en el cual se encuentren. Por ejemplo, el SN

centros docentes

donde en este caso *docentes* cumple la función de adjetivo, ya que complementa al nombre *centros* ; sin embargo, en el SN

docentes buenos

la palabra *docentes* aparece como sustantivo y *buenos* como adjetivo donde se expresa una cualidad de los docentes.
De todas maneras, es un serio problema cuando se maneja una regla que deduzca la categoría gramatical de una palabra en función del "lugar" que esta ocupa dentro de la estructura de SN (Sustantivo Adjetivo), puesto que si se tuviera el SN

buenos docentes

la palabra *docentes* funciona como adjetivo y *buenos* como sustantivo; lo que es sin dudas incorrecto, puesto que uno puede referirse simplemente a *los docentes* pero no a *los buenos*, al faltar un contexto en el que se exprese quienes son esos *buenos* a los que se hace referencia.

Se está en presencia de un caso donde los criterios sintácticos no pueden por sí solos decidir. Este fenómeno se discutirá más adelante cuando se propongan las heurísticas.

Pero sigamos analizando al sustantivo. Los rasgos que comparten con él las palabras que lo acompañan son el género y el número.

Género

En español el género para los sustantivos se divide en **masculino** y **femenino**; no existe el género invariable. Por lo general, el género lo marca el determinante que lo acompaña,

la mesa
el hombre
la mano

¿ Siempre queda claro cuál es el género del sustantivo ? *el docente*
la docente

¿ Qué sucede cuando un determinante ni siquiera está presente en un SN ? *agua*
mano
calle

Este trabajo intenta contestar estas interrogantes cuando se propongan las heurísticas.

También están los casos de:

- nombres epicenos, que son aquellos que no determinan el género:
bebe
persona

- nombres ambiguos, que son los que adoptan los dos géneros pero según el caso tienen distinto significado :
el / la orden

- nombres propios, que tienen por género el sexo a que pertenecen las personas o animales que los llevan.

Ayudan a determinar el género de una palabra, el morfema ***_a*** (femenino) , ***_o*** , ***_e*** (masculino) por citar algunos, aunque existen excepciones como por ejemplo,

día	*cumbre*	*artista*
axioma	*llave*	*testigo*
clima	*noche*	*docente*

Lo importante a destacar es que si bien el género puede deducirse morfológicamente, en general parece ser más conveniente hacerlo con la ayuda o bien de los determinantes,
la llave
el clima

o bien de los adjetivos con los que se combina
noche clara
día lluvioso

Número

En contraste al género, el número gramatical para los sustantivos constituye un sistema más coherente. Puede ser *singular* o *plural*.
La forma singular carece de morfema (en ocasiones se dice que posee morfema 0); mientras que la plural presenta dos morfemas : ***_s*** , ***_es*** ó la ausencia de morfema, donde la forma plural coincide con la singular.

Por lo tanto se puede decir que en general, el número se manifiesta morfológicamente.

Para formar el plural, se agrega al singular de cada nombre alguna de las terminaciones planteadas (si fuera el caso) y los criterios se establecen según la estructura fonológica del singular, la acentuación, etc.
Como se verá más adelante, esto no es tan simple cuando lo que se estudia es la cadena de caracteres; se puede observar que se pueden plantear problemas tales como :
pez *peces*

análisis *análisis*

El trabajo propone un conjunto de heurísticas que permitan decidir el plural de una determinada palabra a partir del singular y vice versa.

Complementos

Los complementos del núcleo dentro de un SN pueden ser :

- adjetivo
- aposición
- complemento prepositivo
- adverbio pronominal
- subordinada adjetiva
- subordinada sustantiva.

En este trabajo, se centra la atención en los adjetivos y en los complementos prepositivos.

- Nombre adjetivo.-

El nombre adjetivo, llamado también simplemente ***adjetivo***, es la parte del SN que se junta al sustantivo para calificarlo o determinarlo. Concuerda con él en género y número. En algunos casos puede ir precedido de un artículo

el buen profesor

De todas maneras nunca se presenta solo (como los sustantivos).

Como se mencionó anteriormente, los sustantivos y los adjetivos tienen características comunes tanto funcionales como formales. Muchos nombres son tanto sustantivos como adjetivos : *docente ,* etc.
Más allá de esto, en muchos casos son fundamentalmente los contextos que hacen referencia a personas los que provocan que los adjetivos sean considerados como sustantivos :

los mejores
los buenos

La función característica del adjetivo es la de atributo del sustantivo. En estos casos se puede colocar o bien inmediatamente después del sustantivo del cual depende :

proceso industrial

o bien junto a un artículo o pronombre que remite anafóricamente al sustantivo del cual depende y lo representa :

la política nacional y la internacional.

Según Bosque en [BOS91] para algunos gramáticos los artículos pasarían a ser "sustantivadores" porque convierten en sustantivo a los adjetivos en los que inciden

el internacional

La propia Real Academia Española considera errónea esta postura y pone el peso en el papel que juega el artículo (o determinante) dentro del sintagma. El tema es que con los sustantivos se expresan determinadas nociones mientras que con los adjetivos se expresan otras.

Wierzbicka en [WIE86] sugiere que los sustantivos "categorizan", determinan clases de objetos, mientras que los adjetivos "describen" propiedades que no constituyen clases. Pertenecer a una clase, significa tener una o varias características que permiten a un individuo de esa clase diferenciarse del resto de los individuos (de otras clases).

Es de destacar, que esta confusión no se presenta en todas las lenguas. Por ejemplo en el inglés, los sustantivos y los adjetivos presentas claras diferencias formales y las gramáticas de estas lenguas las consideran categorías diferentes.
Las clases pueden variar de una lengua a otra. Por ejemplo *young* en inglés y *jeune* en francés son adjetivos y no son sustantivos. Sin embargo en español, *joven* puede funcionar tanto como adjetivo o como sustantivo.

La conversión de un adjetivo en sustantivo obedece muchas veces a la capacidad de ciertas propiedades para pasar de ser descriptores de individuos a denotadores de una clase :

largo (de un vestido)
vestido largo

Los adjetivos se pueden agrupar en tres conjuntos :

- calificativos o especificativos :cumplen una función distintiva, son más denotativos y objetivos y suelen ir detrás del sustantivo.
  Por ejemplo,
  crisis ecológica
  lenguaje natural

- determinativos : para determinar la extensión en que se toma el significado del sustantivo y por lo tanto van antes de este.
  Por ejemplo,
  algún profesor
  mucho calor

- explicativos : dan una opinión del hablante, tienen un carácter más subjetivo y pueden ir o bien antepuestos a los sustantivos.
  Por ejemplo,
  preciosa canción
  estupenda velada

  o bien pueden ir detrás del sustantivo :
  canción preciosa
  velada estupenda

A continuación, se comentan brevemente los mismos rasgos que se trataron para los sustantivos.

Género

De acuerdo a este rasgo, el género para los adjetivos podrían clasificarse en tres grupos:

- invariables : poseen una única forma tanto para el masculino como para el femenino. En general son aquellas palabras que pueden funcionar también como sustantivos; como por ejemplo,

  homicida
  israelí

  También aquellas terminadas en:

  _e *alegre*
  _ble *amable*
  _ente *independiente*
  _al *internacional, social*
  _ar *militar*
  _iz *feliz*
  _or *superior*

  por citar algunos ejemplos.

- femeninos : terminados en ***_a*** y masculinos : terminados en ***_o***
  masculinos Este es el grupo más numeroso y existe la regla que cada uno presupone la existencia del otro; vale decir, que se puede inferir uno a partir del otro

  ecológica *ecológico*
  químico *química*

- no contemplados en los anteriores, es decir, aquellos para los que no existe una regla femenino/masculino genérica

  haragán *haragana*
  receptor *receptora*
  inglés *inglesa*

  y hay que estudiar un poco más detalladamente estos casos.

En el trabajo, si bien se determina el género para cada adjetivo, no se plantea ninguna heurística que permita hallar el femenino (o masculino) correspondiente, a no ser para el caso del segundo grupo.

Número

El tratamiento de los morfemas para el plural es en general común tanto para los sustantivos como para los adjetivos así que no es presentado nuevamente. (Ver el punto correspondiente al Número en el estudio del Nombre Sustantivo).

- <u>Complementos prepositivos.-</u>

Antes de entrar a analizar este tipo de complemento del núcleo de los SSNN, se habla de la preposición y de la conjunción, que funcionan como elementos de enlace.

La gramática tradicional señalaba concretamente, que las preposiciones tienen término y no que los términos tienen preposiciones. Más que términos, de lo que se habla es de <u>complementos de preposición</u>.

Estos se pueden dividir en:

- <u>especificativos</u>, donde el complemento suele tener un carácter más informativo y objetivo, por ejemplo : *madre de Juan*

- explicativos, que son más subjetivos y en general se usan por una razón de estética más que para ampliar el concepto, por ejemplo : *ramas de los árboles*

Es importante destacar, que tanto la preposición como la conjunción son esenciales en la sintaxis, porque si bien no remiten a conceptos o ideas, permiten establecer relaciones puramente gramaticales.

- Preposición.-

Como ya se mencionó, la ***preposición*** es una parte de la oración que no tiene valor por sí sola en el habla, sino que funciona como elemento de relación cuyo significado depende de las palabras por ella relacionadas

anillo ***de*** *oro* (materia)
casa ***de*** *Pedro* (propiedad)

La preposición siempre precede a su complemento (o término), formando con él una unidad sintáctica (y también fonética).
Pueden intercalarse entre la preposición y el término, un determinante

por ***la*** *calle*
en ***el*** *poder*

Unicamente se contraen las preposiciones ***de*** y ***a*** con el artículo ***el***

ganancias ***del*** *capital*
materiales resistentes ***al*** *calor*

De los dos términos relacionados por la preposición, el primero de ellos puede ser un sustantivo, adjetivo, verbo, pronombre, adverbio y se dice que rige a las preposiciones; pero el segundo ha de ser siempre un sustantivo o un equivalente en significación (pronombre, infinitivo, SN, adverbios de lugar o de tiempo).
Además, en ciertos casos la preposición junto con el segundo término forman un "todo lógico" y por lo tanto puede ir precedido de otra preposición, dándose el caso de que vayan dos preposiciones seguidas

de a cuatro
hasta con la madre

Estas "unidades" pertenecen al grupo de las llamadas locuciones prepositivas. Son ejemplos, ***para con , delante de , después de , en lugar de , a fin de*** ,etc.; donde se ve que en algunos casos no se trata de dos preposiciones seguidas, como el caso de *después de*, donde *después* no es una preposición; o que entre dos preposiciones aparece un sustantivo como es el caso de *a fin de*.
No serán tratadas con más detalle, por haberlas dejado en primera instancia fuera de este trabajo.

En resumen, se puede decir que las preposiciones son palabras que encabezan un complemento nominal de otra palabra y lo subordinan a ella.

La lista de las preposiciones es la siguiente [RAE31]:

a	***ante***	***bajo***	***cabe***	***con***	***contra***
de	***desde***	***en***	***entre***	***hacia***	***hasta***
para	***por***	***pro***	***según***	***sin***	***so***
sobre	***tras***				

- Conjunción.-

A diferencia de la preposición que subordina siempre a su término, la ***conjunción*** se limita a coordinar elementos sintácticos de la misma clase no pudiéndose subordinarse unos a otros.
Las más usadas son ***y*** para SN afirmativos y ***ni*** para negativos. La conjunción *y* toma la forma ***e*** cuando precede a palabras comenzadas en **i**_ o **hi**_ (a no ser que el sonido de **i** forme por ejemplo diptongo **ie**).

Este es el tipo de conjunción (coordinación copulativa) "pura", es decir, enlaza términos homogéneos en su función gramatical,

analgésicos y antiepiréticos
oferta y demanda

aunque existen otros estudios que no serán tratados en este documento; a saber, coordinación disyuntiva (*o , u*), adversativa (*más , pero*) , etc.

Con esto se da por concluida la reseña acerca de algunos aspectos de la morfología y también de la sintaxis del español a los solos efectos de ubicar al lector en el contexto del trabajo.

4 ALGUNOS CONCEPTOS DE INFORMATICA DOCUMENTAL

4.1 Introducción

En el tratamiento de sistemas de interrogación en lenguaje natural, el usuario tiene dos tareas fundamentales:

- formular su consulta
- seleccionar y recuperar los documentos adecuados

Se debe tener idea de cuáles son las vías de acceso a la información:

- texto completo (full-text)
- palabras clave
- códigos
- descriptores, etc.

Es decir, cuales son los datos y como se almacenan de manera de facilitar su recuperación a posteriori. En tal sentido, un concepto clave es el de indización, es decir, la forma de describir o caracterizar un documento con la ayuda de los conceptos contenidos en él. Esta indización se puede hacer o bien por texto completo (full-text), donde la consulta se hace comparando todo el texto; o bien por temas o palabras clave dentro de un determinado texto; o bien por descriptores, que son términos que pueden contener más de una palabra y que referencian a un concepto. Los descriptores constituyen la unidad léxica de un lenguaje documental llamado tesauro.

4.2 El Tesauro

Básicamente, un ***tesauro*** (del griego : *thesauri*) consiste de un conjunto de descriptores (conceptos) normalizados , agrupados en familias semánticas y relacionados entre sí.

Es un tipo de lenguaje documental controlado, es decir un lenguaje artificial que acota un sector del lenguaje natural y lo codifica en base a términos preestablecidos por expertos en el área que expresen el contenido de los documentos.
Dentro de los lenguajes controlados, un factor importante que los distingue es el nivel de coordinación entre los términos, ya que para expresar un determinado concepto puede ser necesario combinar varias palabras. En este sentido se puede hablar de lenguajes:

- pre-coordinados : los términos de indización son estructurados, conocidos de antemano, las palabras que los conforman son combinadas en el propio lenguaje del sistema y por lo tanto ya están combinadas en el tesauro.

- post-coordinados : los términos son elegidos a la hora de indizar o buscar un determinado documento. Proporcionan un mayor número de vías de acceso a los documentos.

Un término de un lenguaje pre-coordinado puede consistir de varios términos de un lenguaje post-coordinado vinculados por ciertas relaciones que permitan evitar la ambigüedad.
Un menor nivel de coordinación permite por un lado, una mayor libertad en la asignación en los términos de indización, pero da la posibilidad de recuperar documentos no pertinentes (ruido).

Un vocabulario controlado, consiste en una lista de descriptores normalizados establecida previa al análisis. Define todos los términos que se pueden usar para representar el contenido de un documento. Es una "inversión" inicial importante pero permite lograr mayor rapidez y eficacia en la fase de recuperación de documentos.

En el área de la informática documental y de acuerdo a la norma ISO 2788 - 1974, desde un punto de vista *estructural*, se denomina ***tesauro***, a un lenguaje (vocabulario)

controlado y dinámico de términos agrupados en familias semánticas, donde dichos términos están relacionados jerárquica y asociativamente. Estas relaciones, limitan en algún sentido el significado de los términos; no obstante permiten acceder a los conceptos adecuados para representar el contenido de un documento de una manera más fácil.

Efectivamente, tanto los autores de los documentos y los bibliotecólogos, como los usuarios de los centros de documentación y bibliotecas, utilizan términos diferentes para designar los mismos conceptos. Entonces, desde un punto de vista *funcional*, un ***tesauro*** es un instrumento de control terminológico que sirve para "traducir" en un lenguaje más riguroso el lenguaje natural usado, por un lado por los documentalistas que son quienes registran los documentos y por otro lado por quienes realizan las consultas.

Elementos estructurales

El tesauro consta de dos elementos básicos :

- unidades léxicas :
  - descriptores
  - términos equivalentes
  - identificadores
- relaciones :
  - definitoria
  - equivalencia
  - alternativa
  - jerárquica
  - asociativa

Unidades léxicas.-

Descriptores

Los ***descriptores***, son términos simples o compuestos que expresan un concepto y que han sido designados como *preferenciales.*
Si bien los descriptores están compuestos por una o más palabras, en lo posible su extensión es pequeña. Desde el punto de vista de la estructura gramatical, en general se trata de ***sintagmas nominales***, aunque existen algunas excepciones que serán comentadas oportunamente.

Vale la pena detenerse un instante en este punto ya que es clave para el resto del documento porque ésta es justamente la estructura que presentan las entradas del tesauro.
Los ejemplos que se muestran a partir de aquí, incluyen una serie de símbolos que pertenecen al Tesauro Spines y siguen la Norma ISO 2788 - 1974.

He aquí un ejemplo de descriptor,

2851 FIBRAS DE VIDRIO 29
EN GLASS FIBRES
FR FIBRES DE VERRE
= LANA DE VIDRIO
< FIBRAS INORGANICAS
< FIBRAS NATURALES
< FIBRAS
- INDUSTRIA DEL VIDRIO
- MATERIALES DE CONSTRUCCION
- PLASTICOS REFORZADOS CON FIBRA DE VIDRIO

Notar que estas entradas no solo consisten en sintagmas nominales, sino que también aparecen números y caracteres especiales.

En el ejemplo anterior, se indica que el descriptor FIBRAS DE VIDRIO está vinculado a otros descriptores, como ser FIBRAS INORGANICAS, INDUSTRIA DEL VIDRIO, etc. relacionándose con estos de diferente manera.

Términos equivalentes

Los términos equivalentes o llamados también ***no descriptores***, son términos que representan conceptos no utilizados ni en la indización ni en la recuperación. Se incorporan como una entrada más al tesauro, y es la forma de introducir sinónimos para los descriptores. Su función es simplemente reenviar al descriptor adecuado.

Un ejemplo de no descriptor,

MATERIALES NO IMPRESOS 17
: GRABACIONES LEGIBLES POR COMPUTADOR
: MAPAS
: MATERIALES AUDIOVISUALES

En este ejemplo se tiene el no descriptor MATERIALES NO IMPRESOS mientras que

GRABACIONES LEGIBLES POR COMPUTADOR
MAPAS
MATERIALES AUDIOVISUALES

son descriptores.
Esto significa que si en una consulta aparece MATERIALES NO IMPRESOS , los documentos que involucran tal concepto estarán indizados por alguno de los descriptores colocados debajo.

Es importante notar que tanto en los descriptores como en los no descriptores, no se coloca ningún tipo de signo de puntuación y para el caso de ambigüedad, se limita su significado mediante un término entre paréntesis (el cual también forma parte del no descriptor); como por ejemplo,

GRANOS (ALIMENTOS) 13
=> CEREALES

Entonces, si aparece GRANOS se reenvía al descriptor CEREALES.

En el primer ejemplo elige de tres posibles, en el segundo ejemplo solo hay una posibilidad.
La sintaxis de estos dos ejemplos será explicada más adelante.

Identificadores

Los identificadores consisten en términos que indican nombres de lugares, fechas, fórmulas, siglas, etc.; en general son nombres propios. En algunos tesauros, estos también son descriptores, mientras que en otros conforman una lista separada.

En este trabajo, se empleó el Tesauro Spines, que tiene a los identificadores como un descriptor (o no descriptor) más. Es de destacar que en estos casos, las heurísticas planteadas ocasionan problemas, puesto que no podrán formar el plural de JUAN o el femenino de UNESCO, por citar un par de ejemplos.
Esto se puede solucionar si se tiene a priori la lista de nombres propios y siglas incluidos en el tesauro, cosa con la que no se contó.

Relaciones.-

Las ***relaciones*** entre las unidades léxicas del tesauro son *recíprocas*, es decir, se establecen entre pares de términos y se indican en ambos.
Se pueden clasificar en :

Definitorias :

donde en realidad, esta no es exactamente una relación entre dos términos, sino que consiste en la introducción de una breve explicación del término a los efectos de facilitar su uso. El operador es **SN** : Scope Note (Nota de alcance); y además de no formar parte del descriptor, puede tener diversos usos:

- desarrollar una abreviatura o acrónimo, por ejemplo

7374 UPU 33
....
SN Unión Postal Universal
.....

- excluir algún sentido del descriptor, por ejemplo

1956 FUNCIONES ECONOMICAS 05
.....
SN Relaciones entre variables económicas
.....

- aportar una breve definición, por ejemplo

3397 INFORMATICA 17
.....
SN Ciencia que estudia los medios (físicos y lógicos) para el tratamiento de la información.
.....

Es de destacar, que este tipo de relaciones sólo existe asociada a los descriptores y no está presente en todos.

También se pueden agregar al descriptor sin formar parte de él, una traducción a otros idiomas. En la versión en español del Spines aparece

2986 ANALISIS ARMONICO 20
EN HARMONIC ANALYSIS (EN : ENglish)
FR ANALYSE HARMONIQUE (FR : FRench)
.....

que sirven para asociar determinados términos en las diferentes lenguas en que se edita el tesauro.

Equivalencia :

relación recíproca entre un no descriptor con el descriptor al cual reenvía (que es el que debe ser usado en la inidización o consulta). Vinculan entre si términos que expresan un mismo concepto, que pueden ser considerados como equivalentes y ser tratados como sinónimos en el lenguaje del sistema aunque no sean estrictamente sinónimos.
Por ejemplo, que un no descriptor en donde aparece el descriptor al cual reenvía

FIBRAS MINERALES 29
=> FIBRAS INORGANICAS

Por otro lado existe el descriptor asociado, el cual tiene

3443 FIBRAS INORGANICAS 29
EN INORGANIC FIBRES
FR FIBRES INORGANIQUES
= FIBRAS MINERALES

Alternativas :

son aquellas en las que se asocia no descriptores con más de un descriptor, lo que significa que se debe elegir de entre uno de ellos tanto para la indización como para la consulta. Por ejemplo, se tiene

MATERIALES NO IMPRESOS 17
: GRABACIONES LEGIBLES POR COMPUTADOR
: MAPAS
: MATERIALES AUDIOVISUALES

y por ejemplo en uno de los descriptores aparece,

561 MATERIALES AUDIOVISUALES 17
EN AUDIOVISUAL MATERIALS
FR MATERIEL AUDIOVISUEL
* MATERIALES NO IMPRESOS
.....

Jerárquicas :

son relaciones para indicar niveles de subordinación o superioridad entre términos. Pueden ser de dos tipos :

- genéricas (género / especie) : relaciona un descriptor más general , denominado *descriptor primario* con uno más específico; por ejemplo,

1414 PROGRAMAS DE CONTROL 18
....
< PROGRAMAS DE COMPUTADOR
< SOPORTE LOGICO DE COMPUTADORES
....

En este caso se presentan dos niveles de jerarquía.

- partitivas : relaciona un descriptor con otro que forma parte de él (física o conceptualmente); por ejemplo,

1310 PROGRAMAS DE COMPUTADOR 18
....
< SOPORTE LOGICO DE COMPUTADORES
....
> PROGRAMAS DE CONTROL

Asociativas :

o también llamada de afinidad, se usan para establecer relaciones horizontales, de proximidad conceptual diferente a las anteriores; por ejemplo, una relación entre antónimos, una relación de causalidad, etc.

Por ejemplo,

6238 AGENTES SELLANTES 23
....

- ADHESIVOS
- IMPERMEABILIZACION
- PINTURAS Y BARNICES
- PRODUCTOS QUIMICOS PARA EL HOGAR

Areas temáticas.-

Las entradas al tesauro, están agrupadas en conjuntos de términos que componen una misma área conceptual. En el caso del Spines, los términos se asocian en torno a ciertas ***áreas temáticas***, las que permiten situar a dichos términos en un contexto semántico. Son identificadas con un número y forman parte del término en cuestión.
Es de destacar, que un mismo término puede pertenecer a más de un área temática.

Por ejemplo,

2849 GLACIARES 14 31
.....

Significa que el descriptor pertenece a las áreas temáticas 14 y 31.

Por otra parte, un tesauro puede estar organizado en lo que se denomina ***facetas***, las que también son relaciones semánticas. Estas son identificadas con un número que precede al descriptor (sólo aparece en este tipo de unidad léxica) y forman parte de él.
En el ejemplo anterior, el número 2849 es el código para identificar al descriptor.

Elementos funcionales

Como se mencionó antes en el documento, se ve al tesauro no solo desde un punto de vista estructural, sino también desde un punto de vista funcional como instrumento de control terminológico para indizaciones y/o consultas en lenguaje natural.

¿ Qué se entiende por eficacia en la recuperación de la información en un sistema documental ?
Existen dos aspectos que la caracterizan :

- elementos de exaustividad; son tendientes a disminuir el ***silencio*** (documentos relevantes no recuperados). Para esto se toman algunas normativas, como por ejemplo, emplear solo mayúsculas, no usar acentos, reducir el número de nexos, usar la forma sustantivada, etc.

- elementos de precisión; son tendientes a disminuir el ***ruido*** (recuperación de documentos no pertinentes). Para esto se trata de lograr un buen nivel de pre-coordinación en el lenguaje, una relación entre los términos más rigurosa, agregar ponderación (valor, peso) a cada descriptor en función de la importancia del concepto dentro del documento, etc.

4.3 Algunas características del Tesauro Spines

4.3.1 Introducción

Como se mencionó anteriormente, el Tesauro Spines es la fuente de entrada de información para la prueba de las heurísticas y por lo tanto resulta imprescindible comentar algunas de sus características.

Fue elaborado en el marco de un programa de la UNESCO a partir de 1972, que tenía por finalidad facilitar el manejo e intercambio internacional de documentos y datos de carácter científico y tecnológico tanto a nivel gubernamental como de instituciones de investigación.
Si bien fue concebido en un principio como un sistema de intercambio de información, esto no sucedió como fue previsto y es así que hoy en día su uso es diverso; como por ejemplo, en el tratamiento de documentación científico y tecnológica, consultas a bases de datos informatizadas, en preparación de boletines bibliográficos, en sistemas de clasificación de disciplinas científicas, etc.

El Tesauro Spines no es un diccionario ni una enciclopedia, sino una herramienta (lenguaje) documental que permite lograr la coincidencia en los distintos lenguajes empleados tanto por quienes emiten la información como por quienes la reciben de forma tal de obtener un rápido acceso a los documentos pertinentes en una base de datos.

Por otro lado, su carácter multilingüe facilita el intercambio de información entre personas de diferentes comunidades lingüísticas.

Elaborado entre 1972 y 1975, fue en 1976 que aparece la primera versión para la lengua inglesa. De todas formas, este tipo de obras tiene un carácter dinámico y nunca está terminada definitivamente. A partir de esa primera versión, se ha estado incrementando no solo la cantidad de términos (cerca de 1000 en 1983) sino también agregar nuevas relaciones (en un principio fueron 4000) que incrementan la riqueza estructural y su utilidad.
Hoy en día, el tesauro Spines consta de más de 10.000 términos (entre descriptores y no descriptores) y alrededor de 77.000 relaciones semánticas.

La preparación de versiones para diferentes lenguas no es solo un trabajo de traducción debido a la complejidad de los problemas conceptuales y lingüísticos que requieren una adaptación terminológica. Es aquí donde los códigos numéricos, facetas, asignados a cada descriptor y por lo tanto a cada concepto tienen especial importancia puesto que permiten obtener una concordancia unívoca entre todas la versiones lingüísticas.

4.3.2 Contenido

Se emplea el singular para los términos que representan propiedades específicas, procedimientos particulares, materiales específicos, nombres propios, conceptos que no pueden ser numerados, etc., como por ejemplo

RADIACTIVIDAD
COBRE
JULIO
REASIGNACION

Se emplea el plural para los términos que representan procedimientos y materiales genéricos, objetos y entidades, conceptos que pueden ser numerados, etc., como por ejemplo

LINEAS AEREAS
MATERIALES MAGNETICOS
ESTRELLAS

Los términos se escriben con letras mayúsculas; no tienen un largo superior a 40 caracteres (incluidos los separadores) ; solo se emplean, además de letras, símbolos tales

como #, (,), +, y - ; abreviaturas tales como I+D , C+T , etc. y algunas que indican que los términos se mantienen en el idioma original ING , AL , FR , URSS, etc.

El contenido conceptual del tesauro se divide en 34 áreas temáticas. A cada una de ellas le corresponde un código numérico que figura detrás de cada término. Es de destacar, como ya se dijo que un término puede pertenecer a varios grupos; dentro de cada grupo el orden es alfabético.

A continuación se presenta la estructura de los descriptores y de los no descriptores soportados por el Spines explicando previamente la simbología empleada

Equivalencia
- => **ú sese** todo no descriptor es seguido de un reenvío al descriptor a usarse en la indización
- = **usado por** marca los descriptores escogidos con preferencia a otro u otros; del lado derecho se colocan no descriptores

Alternativa
- : **vé ase** de un no descriptor se reenvía a varios descriptores de los cuales habrá que elegir uno
- * **viene de** reenvío en sentido inverso

Jerárquica
- < Término más genérico
- > Término más específico

Asociativa
- \- Relación entre descriptores

XXXX	Código numérico que indica la faceta (hasta 4 dígitos)
XX	Código numérico para el área temática (puede ser más de uno)

Estructura de los *Descriptores*

XXXX **DESCRIPTOR** en español **XX**
EN Descriptor en inglés
FR Descriptor en francés

Nota de Alcance (si existe, va en cursivas y no se coloca el operador SN)

***** No descriptor
= No descriptor

< Descriptores más genéricos
< (hasta 3 niveles)
<

> Descriptores más específicos
> (hasta 3 niveles)
>

\- Términos relacionados

Estructura de los *No Descriptores*

Con relación de equivalencia : NO DESCRIPTOR XX

Posible Nota de Alcance

=> Descriptor

Con relación alternativa : NO DESCRIPTOR XX

Posible Nota de Alcance

: Descriptor_1
: ....
: Descriptor_n

4.3.3 Ejemplos

A continuación se presentan ejemplos de términos que se pueden encontrar en el Spines.

Ejemplo 1.- Descriptor genérico de primer rango

2007 CENTROS DOCENTES 10
EN EDUCATIONAL INSTITUTIONS
FR ESTABLISSEMENTS D'ENSEIGEMENT
= ESCUELAS
* ORGANIZACIONES (INSTITUCIONES)
> CENTROS DE ENSEÑANZA SUPERIOR
> ACADEMIAS DE CULTURA GENERAL
> COLEGIOS MAYORES
> ESCUELAS DE FORMACION PROFESIONAL
> ESCUELAS NORMALES
> ESCUELAS SUPERIORES DE AGRICULTURA
> ESCUELAS TECNICAS SUPERIORES
> UNIDADES DE ENSEÑANZA E INVESTIGACION
> UNIVERSIDADES
> DEPARTAMENTOS UNIVERSITARIOS
> INSTITUTOS UNIVERSITARIOS DE TECNOLOGIA
> UNIVERSIDADES A DISTANCIA
> UNIVERSIDADES INTERNACIONALES
> UNIVERSIDADES POLITECNICAS
> CENTROS DE FORMACION
> ESCUELAS PRIMARIAS
> ESCUELAS SECUNDARIAS
> INTERNADOS
> ESCUELAS TECNICAS NO SUPERIORES
- ASOCIACIONES EDUCATIVAS
- CENTROS DOCENTES (EDIFICIOS)
- DEFICIENCIA EDUCATIVA
- EDUCACION
- ENSEÑANZA PUBLICA
- ESTADISTICAS EDUCATIVAS
- GESTION DE LA ENSEÑANZA
- INSTALACIONES DOCENTES
- LEGISLACION EDUCATIVA
- MINISTERIO DE EDUCACION
- POLITICA EDUCATIVA
- PROFESORES
- RECURSOS INSTITUCIONALES
- SECTOR DE LA EDUCACION

- SECTOR PRIVADO
- SERVICIOS DE EXTENSION
- TITULOS Y DIPLOMAS

Ejemplo 2.- Descriptor de rango jerárquico medio o de segundo nivel.
Ver que este descriptor aparece en el Ejemplo 1.

7355 UNIVERSIDADES 10
EN UNIVERSITIES
FR UNIVERSITES
= ESCUELAS UNIVERSITARIAS
= FACULTADES UNIVERSITARIAS
< CENTROS DE ENSEÑANZA SUPERIOR
< CENTROS DOCENTES
> DEPARTAMENTOS UNIVERSITARIOS
> INSTITUTOS UNIVERSITARIOS DE TECNOLOGIA
> UNIVERSIDADES A DISTANCIA
> UNIVERSIDADES INTERNACIONALES
> UNIVERSIDADES POLITECNICAS
- ACADEMIAS DE CULTURA GENERAL
- BIBLIOTECAS UNIVERSITARIAS
- CENTROS DE I+D
- CENTROS DE I+D AFILIADOS
- CURSOS NOCTURNOS
- ENSEÑANZA SUPERIOR
- PROFESORES DE ENSEÑANZA SUPERIOR
- SECTOR DE LA EDUCACION SUPERIOR
- TITULACION UNIVERSITARIA
- UNIVERSIDADES (EDIFICIOS)

Ejemplo 3.- Descriptor específico de último rango, o de tercer nivel jerárquico.
Ver que aparece en relación tanto con el descriptor del Ejemplo 1 como con el del Ejemplo 2.

5395 UNIVERSIDADES POLITECNICAS 10
EN POLYTECHNICS
FR UNIVERSITES POLYTECHNIQUES
< UNIVERSIDADES
< CENTROS DE ENSEÑANZA SUPERIOR
< CENTROS DOCENTES
- ENSEÑANZA POLITECNICA
- INSTITUTOS UNIVERSITARIOS DE TECNOLOGIA

Ejemplo 4.- No descriptor con la relación de equivalencia

ESCUELAS UNIVERSITARIAS 10
=> UNIVERSIDADES

Ejemplo 5.- No descriptor con la relación alternativa

ORGANIZACIONES (INSTITUCIONES) 01 04 06 07 08 10 12 15 33
: CENTROS DE I+D
: CENTROS DOCENTES
: EMPRESAS
: GOBIERNO
: GRUPOS
: INSTITUCIONES FINANCIERAS
: ORGANIZACIONES INTERNACIONALES
: ORGANOS RECTORES DE PCT
: RECURSOS INSTITUCIONALES
: SISTEMAS ESTRUCTURADOS

5 ANÁLISIS Y PROPUESTA DE TRABAJO

5.1 Introducción

El objetivo del trabajo es la alimentación de un diccionario con datos morfosintácticos a partir de descriptores de un tesauro. Se realiza un análisis de las estructuras de las entradas - se tomó para la aplicación el Tesauro Spines de la UNESCO [SPN84] - y se propone un conjunto de heurísticas que infieren categoría, número, género y que a continuación se pasan a explicar.

Una vez analizadas las reglas, se verá que se pueden presentar conflictos cuando sea posible aplicar en un caso determinado más de una regla. Estos se estudian al final de la sección donde se establece el criterio adoptado para la elección de la regla.

5.2 Heurísticas

Las reglas llevan un número a los efectos de identificarlas y hacer referencia a ellas en el documento.

Se consideró como punto central, el estudio realizado sobre la morfología de las palabras que componen un SN y se plantea un conjunto de reglas para determinar para cada palabra :

- categória gramatical
- nú mero
- gé nero

En este sentido es que se agrupan las reglas.

Categoría gramatical

Como ya se dijo antes, el trabajo se basa en el estudio de SSNN que conforman las entradas del Tesauro Spines. Como se recordará, existe un diccionario mínimo de palabras "vacías" (conjunto base) tales como preposiciones, artículos, etc. Del conjunto de reglas que se presentan, se deduce el resto de los componentes de la estructura de un SN : Sustantivos y Adjetivos.

Cabe destacar que no se tomaron las frases relativas por ser estas estructuras no significativas en las entradas del tesauro.

Regla 1.-

La primer palabra de todo término es o bien un Sustantivo o bien una Preposición

Es decir que siempre se infiere la categoría gramatical a partir de la forma sintáctica del término.

Se observó que la mayor parte de los descriptores y no descriptores comienzan por un sustantivo o una preposición; rara vez con adjetivo, como

PRIMER MINISTRO

y nunca con un artículo.

Es de destacar, que los términos que comienzan con preposición también son escasos, como por ejemplo el descriptor

EN CURSO

Regla 2.-

Cuando aparecen en un término dos palabras consecutivas (diferentes de las del conjunto base), la primera de ellas es un Sustantivo y la segunda un Adjetivo

Esta regla es extremadamente fuerte puesto que al igual que la Regla 1 limita las estructuras de SN al imponer la categoría a partir de la posición en dicha estructura, por ejemplo

FISICA ESTELAR
MUESTRAS CIENTIFICAS
FUNCIONES ARMONICAS

De todas maneras, en las pruebas aparecen errores del estilo

COMPUTADORES DE LA PRIMERA GENERACION

donde según esta regla, PRIMERA es tomado como sustantivo, cosa que es incorrecto al ser un adjetivo numeral.

Regla 3.-

Si un término consta de 3 palabras distintas de las del conjunto base, la primera de ellas es un Sustantivo y tanto la segunda como la tercera son Adjetivos

La estructura planteada es

PALABRA PALABRA PALABRA

Los dos adjetivos actúan como complementos del sustantivo y el segundo de ellos cumple además una función especificadora

PRODUCTOS QUIMICOS ORGANICOS
NIVEL REGIONAL INTRANACIONAL

Regla 4.-

Un término formado por dos palabras separadas por una Preposición tienen la siguiente estructura :

SUSTANTIVO PREP SUSTANTIVO

Son ejemplos de esta regla,

PROCESO DE DATOS
PROGRAMAS DE COMPUTADOR

Además, la categoría gramatical de la segunda palabra no varía si entre ésta y la preposición se coloca un artículo

SUSTANTIVO PREP DET SUSTANTIVO

Por ejemplo,

TECNOLOGIA DE LA SEGURIDAD
RECICLADO DEL AGUA

Regla 5.-

Luego del adverbio de negación NO, siempre viene un Adjetivo

Este adjetivo actúa como complemento del sustantivo del término en cuestión.

BEBIDAS NO ALCOHOLICAS
ESCUELAS TECNICAS NO SUPERIORES

En los casos que el adverbio esté precedido por un grupo preposicional

[PIEZAS MOLDEADAS DE [[METALES] NO FERREOS]]

donde el adjetivo FERREOS actúa como complemento del sustantivo METALES de dicho grupo preposicional y junto con éste actúan como complemento del sustantivo del término principal (en este caso PIEZAS).

Regla 6.-

En los términos donde aparecen conjunciones (Y , E), éstas se encuentran de la siguiente manera :

SUSTANTIVO CONJ SUSTANTIVO
ADJETIVO CONJ ADJETIVO

Para el primer caso

[OFERTA Y DEMANDA]
[ANALGESICOS Y ANTIEPIRETICOS]
[ESTUDIO DE [TIEMPOS Y MOVIMIENTOS]]

y si existe un complemento, este va luego del segundo sustantivo y complementa a ambos

[[SANIDAD Y SEGURIDAD] OCUPACIONAL]

Para el segundo caso, la conjunción actúa como complemento del sustantivo que precede a la misma

[ARMAS [QUIMICAS Y BIOLOGICAS]]

Regla 7.-

Existe un conjunto de entradas en el tesauro que contienen caracteres especiales tales como guiones, dígitos, etc. Para las palabras que los contienen no se proponen rasgos, a excepción de las palabras contempladas en la Regla 8

Si bien no se analizan, pueden intervenir en algún término cumpliendo una función "complementizadora", es decir, luego de un sustantivo

ALIMENTOS PRE-ENVASADOS
RELACION CAPITAL-TRABAJO
ELEMENTO 104

Ver que por ejemplo no tiene sentido pluralizar

CAPITAL-TRABAJO → CAPITAL-TRABAJOS **

ni hallar el género para los números, etc.

Regla 8.-

Existe un conjunto de entradas en el tesauro que contienen la estructura I+D ó C+T. Estas estructuras son consideradas como Sustantivos

En cualquiera de los dos casos (I+D : Investigación y Desarrollo ; C+T : Ciencia y Tecnología), si son precedidos por una palabra con preposición, o una preposición y determinante, dicha palabra es considerada también sustantivo, ya que se aplican las Reglas 4 y 5

SERVICIOS DE C+T
POLITICA DE I+D DE LAS EMPRESAS

Si una palabra apareciera luego la estructura, dicha palabra es adjetivo por la Regla 2

I+D INDUSTRIAL
I+D INTEGRADA

Si a la estructura la precede o la sigue alguna otra estructura gramatical más compleja, dicha estructura se analiza de acuerdo a otras reglas.

ORGANOS DE FINANCIACION DE C+T
CENTROS DE I+D DE OBJETIVOS MULTIPLES

Número

En la determinación del número (singular / plural) de cada palabra del término analizado, se plantea una serie de heurísticas algunas muy generales, otras que establecen estrategias de singularización basadas en las terminaciones y otras que se basan en un estudio de la concordancia entre núcleo y las palabras que lo rodean en el sintagma.

Las reglas que determinan el número serán fundamentales a la hora de deducir el género, puesto que para inferir éste, se decidió trabajar sobre los singulares de cada palabra y en los casos en que apareciera en un término una palabra en su forma plural, primero se obtiene el singular y luego recién se infiere el género.

Regla 9.-

Las palabras terminadas en _S están en plural y en otro caso están en singular, a excepción de las contempladas en las Reglas 10 y 11

Esta es la regla general para la determinación de los plurales de las palabras.

Regla 10.-

Las palabras terminadas en _IS ó _US son excepción a la regla anterior y presentan la misma forma en singular que en plural

Estas palabras son derivadas del latín.

ANALISIS (MATEMATICAS)
CRISIS ECOLOGICA

Regla 11.-

Las palabras de largo menor o igual a tres terminadas en _S están en su forma singular

Este conjunto de palabras constituye una excepción a la regla general de terminaciones en _S (Regla 9). Son los casos por ejemplo de palabras tales como

MES
GAS

Regla 12.-

En función de determinadas terminaciones, si una palabra está en plural, se calcula su forma singular según tres casos diferentes :

_CES → ***sustituye las tres últimas letras por _Z***

_voc_ES
_voc+LES
_voc+RES
_ASES
_ESES
_D_S
_N_S
→ ***elimina las dos últimas letras***

_AS
_OS
_J_S
_T_S
_U_S
_V_S
_OIDES
_SES
_cons+LES
→ elimina la ultima letra

Cualquier otro caso no contemplado → ***elimina la última letra***

A continuación se muestra el orden de aplicación de la regla:

- si la terminación es _ASES , _ESES , _voc+RES : se eliminan ES

Ejemplos:

CARACTERES	→	CARACTER
TRABAJADORES	→	TRABAJADOR
GASES	→	GAS
MESES	→	MES

- si la terminación es _AS , _OS , _ (J | T | V)[2]+voc+S , _U_S ,

_SES , _OIDES , _cons+LES : se elimina la S

Ejemplos:

ISOPRENOIDES	→	ISOPRENOIDE

[2] Como excepción : *relojes*

IMPUESTOS → IMPUESTO
TRAJES → TRAJE
MULTIPLES → MULTIPLE

- si las tres últimas letras son CES, se cambian por Z[3]
Ejemplo:
PECES → PEZ

- si se tiene la terminación _voc+_ES , _voc+LES : se eliminan ES
Ejemplo:
ARBOLES → ARBOL

- si la terminación es _voc+D+voc+S : se elimina la dos últimas letras
Ejemplo:
VIRTUDES → VIRTUD

- cualquier otro caso, elimina la S (que es la última letra)

Es de destacar que esta regla se fue armando en función del estudio de la morfología realizado previamente y tomando en cuenta los casos encontrados; por lo tanto pueden existir algunos no contemplados. Es por consiguiente claramente empírica.

Regla 13.-

Cuando una palabra está en su forma singular, se calcula su forma plural según tres casos distintos

1)

_Z → ***se sustituye la Z por la cadena CES***

2)

_A
_E
_O
_T
} → ***se les agrega la S al final***

3)

_D
_I
_L
_N
_R
_U
_Y
_AS
_ES
} → ***se les agrega la terminación ES***

4) ***Cualquier otro caso no contemplado considera igual a la forma singular y plural***

A continuación se muestran algunos ejemplos :

Caso 1 :
CAPAZ → CAPACES

[3] Como excepción : *enlaces , vértices*

Caso 2 :
Es claro cuando la última letra es una vocal (exceptuando la U y la I). En el caso de la terminación en la consonante _T

ROBOT → ROBOTS

Caso 3 :
Considera algunos singulares terminados en S. Se trata de palabras de largo menor o igual a 3

AS → ASES
MES → MESES

Otros ejemplos son:

VIRTUD → VIRTUDES
HOGAR → HOGARES
LEY → LEYES

Caso 4 :
Por defecto la regla indica que cualquier otro caso se entiende como que el singular y el plural tienen la misma forma; son ejemplos

PH
LATEX

Esta regla se construyó también en función del estudio y las pruebas realizadas.

Aquí aparece nuevamente el problema del tratamiento de los nombres propios. Cabe acotar que de contar con un diccionario que contenga una lista tanto de los nombres propios, como las siglas, podrían ser excluidos del tema flexiones.
Como las entradas del tesauro están escritas en mayúsculas tampoco está la posibilidad de hacer un análisis de cada palabra en función de la letra con la que empiece (los nombres propios suelen comenzar con mayúsculas).

Regla 14.-

El número de un Complemento (Adjetivo) determina el número del Núcleo (Sustantivo) en estructuras de términos del tipo

SUSTANTIVO ADJETIVO
SUSTANTIVO ADJETIVO ADJETIVO

Los casos citados en la regla son los más representativos (luego existen las variantes, puesto que por ejemplo luego de la primera de las estructuras consideradas en la regla, puede venir algún otro tipo de complemento, por ejemplo una preposición.

En cualquiera de los casos planteados, en función del número del adjetivo (en caso que existan dos se habla del primero de ellos) se infiere el número del sustantivo.

Género

Para la determinación del género correspondiente a cada palabra de un término, no solo se plantea un conjunto de heurísticas que lo deducen a partir de las terminaciones, sino que se presentan reglas para deducir el género de una palabra cuya categoría gramatical es sustantivo en función o bien de otra palabra (adjetivo que lo califica) o bien de un determinante (que lo precede).

Cabe notar que en estos casos se pueden presentar conflictos entre los diferentes criterios a tomar para determinar el género, y es entonces que habría que darle una cierta ***prioridad*** o ***peso*** a las reglas para que se sepa que para una palabra, se infiere el género por determinada regla. Esto permitiría que se tenga información acerca de cual fue la regla por la

cual se infirieron determinados rasgos para una palabra, a los efectos de poder cambiarlos en función del ***peso*** (Ver Capítulo 7, Extensiones).

Regla 15.-

El género se calcula sobre la forma singular

Esta regla parte de la base de que las heurísticas propuestas en las Reglas 9 a 12 funcionan bien; es decir, como ya se comentó anteriormente esta regla habla de calcular el género en base a la forma *singular* de la palabra.

Regla 16.-

Según la categoría gramatical, una palabra puede ser de género : masculino
femenino
invariable

Los dos primeros casos vale tanto para los sustantivos como para los adjetivos, mientras que el tercer caso (género invariable, es decir tanto femenino como masculino), vale solo para los casos de algunos adjetivos que según el contexto cumplen la función sustantivo.

Regla 17.-

Las palabras terminadas en _IS son de género femenino y las terminadas en _US son de género masculino

Esta regla trata a las excepciones comentadas en la Regla 10 respecto a las palabras derivadas del latín. Estas presentan una única forma para el singular y para el plural.
Esta regla expresa el género según cada caso.

Regla 18.-

Las palabras de largo menor o igual a tres y terminadas en _S son de género masculino

Esta regla al igual que la anterior, considera las excepciones de la Regla 11 e infiere son de género masculino. Como se recordará además, estas palabras están en singular. Por ejemplo

GAS

Regla 19.-

Si la palabra es un Adjetivo y su terminación es A, entonces el género es femenino y admite también el masculino terminado en O

Con esta regla, se pretende encontrar el masculino de ciertos adjetivos. Es de destacar lo importante que es el saber previamente no solo la terminación sino también la categoría gramatical de la palabra (debe ser adjetivo), puesto que la regla no vale para los sustantivos, o por lo menos, no siempre y por tanto no se puede generalizar,

Regla 20.-

Si la palabra es un Adjetivo y su terminación es O, entonces el género es masculino y admite también el femenino terminado en A

La idea es similar a la de la regla anterior, valiendo las mismas consideraciones.

Regla 21.-

Si la palabra es un Sustantivo y su terminación es A, entonces el género es femenino

Esta regla funciona prácticamente sin inconvenientes aunque presenta casos de excepción, algunos de los cuales son comentados en la próxima regla.
Como se explicó en la Regla 19, no se construye la correspondiente palabra en masculino puesto que a pesar de tener una raíz común no es posible lograr un mecanismo automático de construcción.
Existe otro tipo de error, en palabras como

DIA
CLIMA

pero en términos generales se puede decir que los resultados son buenos.

También está el caso que por cuestiones fonológicas una palabra sería femenino y sin embargo llevan un artículo masculino

EL AGUA

pero que en su forma plural sí adopta el femenino

LAS AGUAS

Notar que en la mayoría de los sustantivos el género es *gramatical* y no *natural*, solo tiene sentido una forma. No ocurre lo mismo con los adjetivos que como se observó en las reglas anteriores, suelen presentar las dos formas. No obstante, hay ocasiones en que la forma es única y depende de la terminación. Esto se analiza en las Reglas 24 a 27.

Regla 22.-

Si la palabra termina en _EMA entonces es de género masculino

Esta regla es una excepción a la Regla 21.
Son ejemplos de este tipo de casos

MORFEMA
SISTEMA
PROBLEMA

Regla 23.-

Si la palabra es un Sustantivo y su terminación es O, entonces el género es masculino

Al igual que para el caso anterior, la heurística funciona aceptablemente y existen unos pocos casos de error tales como

RADIO

Vale la misma consideración que para la regla anterior en cuanto a que no se puede obtener su correspondiente en el género opuesto; o el caso en que la misma palabra admiten los dos géneros

(EL) RADIO
(LA) RADIO

aunque sus significados en cada caso son distintos.

Regla 24.-

Son de género femenino las palabras terminadas en :

_CION _SION _TION _D

Algunos ejemplos,

CANCION
EROSION
GESTION
SALUD

Regla 25.-

Son de género masculino las palabras terminadas en :

_E _R _T _Z
_AL _EL _IL _OL _UL
_AN _EN _ON (salvo las excepciones de la regla anterior
_CION _SION _TION)

Por citar algunos ejemplos,

PIE	HOGAR	ROBOT	PEZ	
MATERIAL	CARTEL	PRETIL	METANOL	BAUL
PLAN	NEON			

Cabe acotar que en este grupo se encuentran palabras que son la forma infinitiva de los verbos, los cuales son considerados sustantivos masculinos y en singular. Esto traerá problema al querer construir su plural.

Regla 26.-

Son de género invariable los Adjetivos terminados en :

_E _Z _AL _IL

Son ejemplos,

INDUSTRIAL
SUTIL
DOCENTE
CAPAZ

Regla 27.-

Por defecto, se consideran todas aquellas palabras no contempladas en las reglas anteriores como de género masculino

Si bien esta regla es muy fuerte, los resultados obtenidos con las pruebas realizadas, no presentaron mayores dificultades

Regla 28.-

El género de un Complemento (Adjetivo) determina el género del Núcleo (Sustantivo) en estructuras de términos del tipo

SUSTANTIVO ADJETIVO

SUSTANTIVO ADJETIVO ADJETIVO

El orden de aplicación es importante y como la determinación del género se establece en los singulares, esta heurística va a plantear por ejemplo, para la primer estructura :

- si el género de la palabra que ocupa el lugar del adjetivo es femenino, entonces el género de la que ocupa el lugar del sustantivo es femenino también;

- si el género de la palabra que ocupa el lugar del adjetivo es masculino, entonces el género de la que ocupa el lugar del sustantivo es masculino también;

- si el género de la palabra que ocupa el lugar del adjetivo es invariable, entonces se halla el género de la que ocupa el lugar del sustantivo independientemente (reglas 22 y 24 a 27)

Para el caso de la estructura de término con dos adjetivos, primero se chequea el género del adjetivo inmediato posterior al sustantivo calculándose según éste el género de dicho sustantivo y para el caso que es invariable, se hace lo mismo pero con el segundo adjetivo.

Esta regla tiene mayor ***peso*** que aquellas que calculan el género de una palabra en forma aislada, puesto que se está exigiendo concordancia entre el género del sustantivo y el del (o los) adjetivo(s).
Por ejemplo, en

RECONOCIMIENTO OPTICO DE CARACTERES

el adjetivo OPTICO es masculino y por lo tanto RECONOCIMIENTO es masculino
El sustantivo CARACTERES (previo obtención de su forma singular) se calcula con las otras reglas y se infiere masculino por Regla 25.
En el ejemplo

INTELIGENCIA ARTIFICIAL

el adjetivo ARTIFICIAL es invariable, por lo tanto se calcula el del sustantivo INTELIGENCIA con otras reglas, deduciéndose femenino por la Regla 21.
Lo mismo ocurre con el descriptor

NAVES ESPACIALES

Aquí nuevamente, puesto que el género del adjetivo ESPACIAL (singular de ESPACIALES) es invariable, el género del sustantivo NAVE se calcula con la Regla 25 obteniéndose masculino.
Sin embargo, considérese el ejemplo del descriptor

NAVES ESPACIALES TRIPULADAS

como ESPACIAL es invariable, se calcula el género de TRIPULADAS, que es femenino (por Regla 21) y por tanto se infiere que NAVE es de género femenino (concordancia sustantivo - adjetivo).
Este es un caso en donde la palabra NAVE debe modificar sus rasgos si hubiera sido agregada al diccionario en función del descriptor NAVES ESPACIALES.

Regla 29.-

Si una palabra es un Sustantivo precedido por un Determinante, es éste el que determina el género de dicho Sustantivo :

_A o _AS	→	***el Sustantivo es de género femenino***
otro caso	→	***el Sustantivo es de género masculino***

Son ejemplos de terminaciones en A : LA , LAS , UNA , ALGUNA , etc.
Los otros casos abarcan : EL , LO , LOS , UN , UNO ,etc.
Es importante volver a destacar el orden de aplicación de estas heurísticas, pues va a estar el caso del término

RECICLADO DEL AGUA

donde generará que AGUA es de género masculino.
Recordar que este caso fue comentado en la Regla 21.

5.3 Conflictos entre las heurísticas

En este punto se comentan los conflictos que se presentaron a la hora de aplicar las reglas para cada entrada del Tesauro Spines.

Es de destacar que los conflictos se plantean cuando se elige la regla a aplicar al inferir los rasgos de una palabra cuando ésta está asociada con un artículo, con otra palabra, o con más de una palabra.
A continuación se plantean dichos conflictos y se detalla como se solucionaron.

- **Conflictos a nivel de Categoría Gramatical**

Cabe recordar que se tenía la Regla 1 en la que se le asignaba la categoría de ser Sustantivo a la primer palabra de un término siempre y cuando no fuera una Preposición. Esta regla es muy fuerte, puesto que limita la cantidad de posibles conflictos que se puedan presentar. Debido a que en el tesauro manejado los SSNN **no** comienzan con Determinante y los casos en que comienzan con Adjetivo son escasos. Por el contrario, sí es posible que aparezcan adjetivos por delante de sustantivos en algunos términos, lo que constituye sin duda un problema. No obstante ello, el tipo de adjetivos en estas condiciones es limitado (por ejemplo, los adjetivos numerales) y como se menciona más adelante, este tipo de error es posible de solucionar, aunque al momento no es considerado por la implementación (Ver Capítulo 7, Extensiones).

Otro tema, que en realidad no se trata de un conflicto, es el hecho de que en ocasiones, una palabra considerada como sustantivo también puede ser tratada como adjetivo (en otro contexto).

CENTROS DOCENTES
DOCENTE DE INGENIERIA

Como se recordará, cuando se hizo el estudio de la categoría gramatical Sustantivo, se mencionó que existen muchos casos en los cuales una misma palabra puede cumplir tanto la función de sustantivo como la de adjetivo diferentes contextos.
En la implementación se optó simplemente por reconocer las dos acepciones como válidas (lo cual es correcto).
En estos casos, la solución es colocar con dos entradas en el diccionario la misma palabra con distinta categoría gramatical según el caso. Esto no se contradice con el hecho que se mencionaba de modificar los rasgos de una palabra en función del peso de la regla por la cual fueron calculados; puesto que se limita a modificar únicamente el género.

- **Conflictos a nivel de Número**

En cuanto a las operaciones realizadas para determinar el número de una palabra, no existen conflictos entre más de una regla.
Dejando de lado las reglas definidas para la deducción del número, el resto de las mismas trabaja sobre la forma singular, esto es, si la palabra analizada aparece en su forma plural, entonces primero se halla el singular, para luego después hallar el género.

Se pueden presentar conflictos en la aplicación de las reglas que determinan el número de una palabra (cuando ésta es un nombre sustantivo) en función de las palabras que la rodean. En tal sentido, se adoptó el siguiente criterio para determinar el número :

- en función del *determinante* que la precede

- en función del (los) *adjetivo(s)* que la suceden. Si existen dos adjetivos consecutivos que complementan al sustantivo, se considera el primero de ellos y el segundo debe concordar con él en número.

- en función de la terminación, como término aislado (Regla 9).

- **Conflictos a nivel de Género**

Para la obtención del género de las palabras es donde se presentan problemas. Estos surgen a la hora de reconocer la estructura del término, es decir, cuando se elige cuales deben ser las funciones apropiadas para inferir dicho género.

La Regla 28 explicaba cual era el tratamiento en los casos de las estructuras

$PALABRA_1$ $PALABRA_2$
$PALABRA_1$ $PALABRA_2$ $PALABRA_3$

El problema se presenta cuando la estructura de término es la siguiente:

$PALABRA_1$ PREPOSICION DETERMINANTE $PALABRA_2$ $PALABRA_3$

Según lo explicado en la Regla 28, como se tienen dos palabras seguidas $PALABRA_2$ y $PALABRA_3$ (que se corresponden con Sustantivo y Adjetivo respectivamente), se calcula el género de $PALABRA_2$ en función del género de $PALABRA_3$.

Sin embargo, en este otro caso existe un nuevo elemento en la estructura del término que hace que la elección sea otra, el DETERMINANTE.
Como debe existir concordancia en los SSNN entre el Determinante y el Núcleo (DETERMINANTE y $PALABRA_2$), en caso de tener un determinante terminado en A, el género de $PALABRA_2$ (el del sustantivo) es **femenino** y en cualquier otro caso, el género es **masculino**.

Cabe resaltar entonces que se elige la opción de inferir el género de un sustantivo en función del determinante que lo precede ante la opción de calcular el género del sustantivo en función de la del adjetivo que lo sucede.
Se trata por consiguiente de reglas con diferente ***peso***.
Si existe ya una palabra en el diccionario, el rasgo correspondiente al género debe ser pasible de modificación.

Por lo expuesto, a la hora de elegir los ***pesos*** asociados a las heurísticas, el emplear el criterio de que en caso de tener un determinante se calcule el género de la palabra que le sigue en función de él, el peso de la función asociada a estas reglas deberá ser superior a la aplicación de la Regla 28, es decir la inferencia de género por concordancia sustantivo - adjetivo.

Para el caso del cálculo del género de $PALABRA_3$ que es un adjetivo, la deducción de dicho género se hace como el de una palabra sola (sin tener en cuenta la concordancia). Esto es posible de mejorar, puesto que se puede deducir a partir del género del sustantivo (que ya fue calculado), al deber de tener el mismo género por concordancia entre Núcleo y Complementos. Esto se puede ver en el Capítulo 7 cuando se proponen Extensiones al trabajo.

En resumen, para la resolución de los conflictos en relación a la deducción del género de una palabra, se establece un criterio similar al planteado para el tratamiento de los conflictos en el número. Se tiene entonces, que el género de una palabra se infiere:

- en función del *determinante* que la precede

- en función del (los) *adjetivo(s)* que la suceden. Si existen dos adjetivos consecutivos que complementan al sustantivo, se considera el primero de ellos y el segundo debe concordar con él. Si el género del primero fuera invariante, se considera el segundo y la concordancia se hace con éste.

- en función de la terminación, como término aislado.

6 IMPLEMENTACIÓN Y RESULTADOS OBTENIDOS

6.1 Implementación informática

Para la implementación se construyó un parser que toma como entrada una secuencia de descriptores y no descriptores del Tesauro Spines[SPN84] y obtiene para cada palabra de la entrada, los tres rasgos que se han venido tratando hasta el momento (categoría, número y género).
A tales efectos, se realizó una gramática para todas las posibles entradas del tesauro. Se construyó un analizador lexicográfico y un analizador sintáctico que permiten reconocer los elementos de cada término y aplicar las heurísticas propuestas.

La gramática consiste básicamente de las siguientes producciones:

(1) ***entrada : termino***
(2) ***| termino paren***

(3) ***termino : nucleo***
(4) ***| modificador***
(5) ***| nucleo modificador***
(6) ***| conj_term***
(7) ***| PALABRA PALABRA PALABRA***

(8) ***nucleo : PALABRA***
(9) ***| PALABRA PALABRA***

(10) ***modificador : gp***
(11) ***| modificador gp***
(12) ***| NO PALABRA***
(13) ***| gp NO PALABRA***

(14) ***conj_term : nucleo Y nucleo***
(15) ***| nucleo Y nucleo modificador***
(16) ***| nucleo Y modificador***
(17) ***| nucleo modificador Y nucleo***
(18) ***| nucleo modificador Y modificador***
(19) ***| nucleo modificador Y nucleo modificador***

(20) ***termino : PALABRA PAL_GUI_PAL***
| PAL_GUI_PAL
| PAL_GUI_PAL paren
| PALABRA gp PAL_GUI_PAL
| PAL_DIG
| PALABRA PAL_DIG
| NUM_GUI_NUM

(21) ***gp : PREP nucleo***
(22) ***| PREP PREP nucleo***
(23) ***| PREP DET nucleo***
(24) ***| PREP DET NRO***

(25) ***paren : PARABR nucleo PARCIE***
| PARABR nucleo modificador PARCIE
| PARABR modificador PARCIE
| PARABR PALABRA PREP DET PARCIE

Como se podrá notar, se están dejando de lado cuestiones de los descriptores y no descriptores tales como los códigos numéricos; traducción al inglés y francés, relaciones. Acá se pretende mostrar los tipos de estructura presente en el Tesauro Spines. No obstante, todas estas características de las entradas a la hora del reconocimiento de los términos del tesauro, si se tomarán en cuenta.

Con esta salvedad, se analizan las producciones de la gramática presentada arriba.

Cada entrada del tesauro es entonces o bien un término (descriptor o no descriptor) o bien un término seguido de una estructura entre paréntesis, lo cual es representado por las producciones (1) y (2).

Como se explicó cuando se habló de la estructura del tesauro Spines, entre paréntesis se coloca una expresión a los efectos de aclarar el alcance del término analizado, como por ejemplo,

MATRICES (MATEMATICAS)

Otros casos en que los que se usan paréntesis son cuando el concepto que se está analizando es ambiguo.
Lo que se coloca es un especificador al existir más de una interpretación.
Ejemplos de este tipo son

PIEL (ANATOMIA)
PIEL (VESTIDOS)

Esto sucede en aquellos casos para los cuales en español una misma palabra representa dos conceptos diferentes y se usan los paréntesis para el calificativo de los homónimos, quienes también forman parte del término.

Por otra parte puede presentarse el caso de una palabra que es mantenida en su idioma original, y entre paréntesis se coloca una sigla que identifica el país de origen, como por ejemplo

DOM TEKHINIKI (URSS)
INPUT (ING)
LEASING (ING)

Existen otros casos en que la expresión es una estructura que no se trata exactamente de un término (descriptor o no descriptor). Son casos donde entre paréntesis se coloca un complemento especificador del concepto, como por ejemplo

VITAMINA P (COMPLEJO DE LA)

Todas estas estructuras son tomadas en cuenta en las producciones (25).

Luego, cada término puede tener alguna de las estructuras presentadas en las producciones (3) a (7) y (20).

Con ***nucleo*** , se muestra que dicha "variable" puede ser o bien una palabra (cuya categoría gramatical se impone que es un sustantivo) o bien dos palabras consecutivas (que corresponden a Sustantivo Adjetivo respectivamente).
Por otro lado, notar que la producción (7) indica que un término puede ser una secuencia formada por tres palabras lo cual no se consideró como *nucleo*.

Un caso distinto es el de la producción (4), donde aparece ***modificador*** como un posible término del tesauro. Esto es debido a que existen entradas (pocas) que comienzan con una preposición.
Las producciones (5) y (6) indican que un término también puede ser un *nucleo* con un *modificador* o bien una conjunción de términos.

Con ***modificador*** se pretende representar los complementos y estos pueden ser :

(10) : un ***grupo preposicional (gp)*** , que puede ser preposición + núcleo (21) ; preposición + determinante + núcleo (23) ; y dos casos especiales, que son para validar términos tales como

INSTRUMENTOS **DE A** BORDO (22)

donde aparecen dos preposiciones juntas (que son uno de los tipos de locuciones prepositivas), el caso del término

DESPUES DEL 2000 (24)

donde aparece un número luego de una preposición + determinante (recordar el tratamiento para el caso de preposiciones contractivas DEL = DE + EL).

(11) : fundamentalmente para el caso de permitir una secuencia de *gp* y reconocer términos tales como

DISPOSITIVOS DE RECONOCIMIENTO DE CARACTERES
POLITICA DE I+D DE LAS EMPRESAS
PLASTICOS REFORZADOS CON FIBRA DE VIDRIO

(12) : esta producción contempla los casos de

BEBIDAS NO ALCOHOLICAS
ESCUELAS TECNICAS NO SUPERIORES

(13) : en estos casos, la palabra que siga al adverbio de negación NO, es un Adjetivo cuya función es la de complementar al sustantivo del *gp* que lo antecede.
De todas formas, esta observación es a los efectos de tener en cuenta las reglas que se disparen cuando se detecte la presencia del adverbio. Son ejemplos,

PIEZAS MOLDEADAS DE METALES NO FERREOS

Con ***conj_term*** se pretende representar a los términos que aparecen unidos por una conjunción.
Por lo que entonces se analizan :

(14) : contempla varios casos,

i.- la conjunción de dos sustantivos
[HERIDAS Y LESIONES]

ii.- la conjunción de dos sustantivos con un adjetivo que los complementa
[[SANIDAD Y SEGURIDAD] OCUPACIONAL]

iii.- un sustantivo y dos adjetivos unidos por la conjunción y que actúan ambos como complemento de dicho sustantivo
[CIENCIAS [FISICAS Y NATURALES]]

Cabe acotar que se deja de lado el caso de la conjunción entre dos términos ambos con la estructura Sustantivo Adjetivo la cual no se presentó.

(15) : aquí el alcance del *modificador* es la conjunción de la estructura formada por los núcleos, es decir que es

(nucleo Y nucleo) modificador

[[OFERTA Y DEMANDA] DE TRABAJO]

(16) : en este caso el *modificador* actúa sobre el sustantivo del *nucleo*. Un ejemplo de este tipo de término es

[HORTALIZAS [VERDES Y DE HOJAS]]

No es un caso de los comunes, pero se puede presentar. Aquí la palabra que viene luego de la preposición es un sustantivo.

(17) : en este caso, tanto la palabra del *nucleo* que precede al *modificador* como la que forma parte del propio *modificador*, y la del *nucleo* correspondiente al lado derecho de la conjunción son las tres sustantivos; la preposición del *modificador* actúa sobre la conjunción, por lo que el alcance de la estructura es en realidad desarrollándola un poco más

nucleo PREP (nucleo Y nucleo)

Notar que cuando se hizo referencia al *nucleo*, se habló de **una** palabra (sustantivo) y no de dos (estructura Sustantivo Adjetivo). Esto se debió simplemente al hecho que estos casos de núcleos de dos palabras no se presentan. De todas formas, si aparecieran, no existe problema puesto que igualmente son contemplados por el parser.
Un ejemplo donde se aplique esta regla es

[ESTUDIO DE [TIEMPOS Y MOVIMIENTOS]]

(18) : este caso es lo opuesto al anterior, ya que ahora se tiene dos estructuras *gp* unidas por una conjunción y todo esto funciona como complemento de *nucleo*.
La estructura es

nucleo (modificador Y modificador)

Un ejemplo,

[FILOSOFIA [DEL ESPACIO Y DEL TIEMPO]]

donde acá se puede ver que se omite el sustantivo FILOSOFIA delante del segundo *modificador*

(19) : este es el caso más general y no se encontraron ejemplos en la muestra considerada.

Siguiendo con el análisis de términos del tesauro, en las producciones (20), se consideran todas aquellas expresiones que se corresponden con términos que contienen caracteres especiales. Este conjunto de reglas son colocadas para poder abarcar varios de los SSNN posibles que aparecen en el tesauro. Se contemplan por ejemplo, términos tales como :

RELACION CAPITAL-TRABAJO
PL1

Aquí se está dejando de lado las estructuras I+D y C+T, que se corresponden a sustantivos y son contempladas como tales. En la implementación tienen un tratamiento especial al tener forma única para el plural y el singular y no se les asigna género.

Como ya se mencionó en la Sección 4.4 (Algunas características del Tesauro Spines), la estructura de los descriptores y de los no descriptores no es solamente un SN sino que aparece además un conjunto de símbolos que muestra las relaciones entre pares de términos.

La idea es que el reconocedor verifique la estructura de tales términos tal como vienen en el tesauro, es decir, con los códigos numéricos para las facetas y las áreas temáticas, su traducción al inglés y al francés y todos aquellos otros términos de alguna manera relacionados.

En cuanto al lenguaje elegido en la implementación de la heurísticas fue C y se emplearon las herramientas Lex y Yacc para la construcción de los analizadores lexicográfico y sintáctico respectivamente.
En tal sentido, la gramática definida en el Yacc difiere un poco de la expuesta antes en este documento solo a los efectos del eliminar ambigüedades propias del análisis sintáctico.

6.2 Resultados obtenidos

En este punto se comenta cuales fueron los resultados obtenidos en las pruebas realizadas sobre las heurísticas con una muestra de aproximadamente 1100 términos diferentes entre descriptores y no descriptores elegidos en forma aleatoria, lo que constituye un poco más del 10% de la totalidad y que contienen a su vez más de 1200 palabras diferentes.

Con este conjunto de prueba se obtuvo alrededor de 50 errores. A continuación se pasan a detallar alguno de ellos, ya que muchos son posibles de solucionar.

- alrededor de 10 son producto de ser o contener estructuras que no se corresponden a las de los SSNN analizados. Como se recordará, la Regla 1 impone que la primer palabra de un término es o bien un sustantivo o bien una preposición. Justamente estos errores son producto de la excepción a dicha regla, donde un adjetivo aparece delante del Sustantivo que califica, como por ejemplo

  PRIMER MINISTRO
  COMPUTADORES DE LA SEGUNDA GENERACION
  NUEVOS PRODUCTOS

- en la muestra, aparecen 3 palabras en otra lengua,

  DOM TEKHINIKI (URSS)
  INPUT (ING)
  LEASING (ING)

- otros 3 errores fueron por aparición de nombres propios, los cuales son tratados como sustantivos y eso es lo que ocasiona los problemas, como ya se dijo, por ejemplo a la hora de construir la forma plural

  JULIO
  DIOS
  ALGEBRA DE BOOLE

Es importante destacar que en estos casos se consideró que como este tipo de errores (al igual que las palabras en otras lenguas) estaban ya detectados, solo se tomaron a los efectos de comprobar efectivamente los problemas previstos.

- otro tipo de error encontrado es el de los verbos en infinitivo. Como ya se comentó en la Regla 25, estas palabras eran considerados como sustantivo, en masculino y en singular. El error aparece cuando se pretende construir el plural, por ejemplo

  APRENDER → APRENDERES
  ESCRIBIR → ESCRIBIRES

- el resto de los errores son los casos de género mal; éstos son producto de aplicaciones de la deducción del mismo en función de los morfemas derivativos y constituyen en muchos casos excepciones a las reglas 17 a 29.

Es de destacar que este último tipo de error, es posible de solucionar de contar con la información asociada a la palabra de cual fue la regla que se aplicó para obtener sus rasgos. Esto ya fue explicado en la sección 5.3 (Conflictos entre las heurísticas).

En el Apéndice se adjunta además un ejemplo de la ejecución del parser para ver el funcionamiento de las heurísticas con ejemplos de términos del Tesauro Spines.

7 CONCLUSIÓN Y EXTENSIONES

El primer punto importante a resaltar es el hecho de los buenos porcentajes obtenidos a partir de las pruebas realizadas. El éxito fue del orden del **95%** en la determinación de los rasgos de cada palabra, de lo cual se desprende como conclusión primaria la ***buena performance*** de las heurísticas propuestas.
Es importante destacar que dentro de los errores reportados se encontraron siglas y nombres propios producto de elegir términos en forma aleatoria.

Por lo tanto, como conclusión se considera que en términos generales se obtiene un buen porcentaje de éxito. Luego de una etapa inicial de afinamiento de las heurísticas por "*ensayo y error*" se llegó a un buen nivel de generalidad en las reglas de manera de poder verificar contra un mayor número de palabras.

Se planteadas una serie de extensiones al trabajo, donde la más importante es la implementación física del *diccionario* y su integración a un sistema que manipule la información.

Se piensa que una palabra tenga una estructura de registro con las siguientes características,

palabra
nombre
plural
género
cat_gramatical
peso_regla
fin_palabra

donde la palabra aparece en su forma singular (en *nombre*) y en *peso_regla* un identificador de regla que indica la prioridad asociada a la regla empleada para identificar rasgos, más concretamente el género, y que es posible alterar en aquel caso en que una palabra pudiera variar en su género según la heurística por la cual fue determinado.
Recordar que una palabra puede aparecer duplicada en el diccionario, son los casos en que cumpla la función de sustantivo ó adjetivo dependiendo del contexto.

El diccionario es una lista de palabras enlazadas por orden alfabético,

diccionario
pal : palabra
punt : ↑*diccionario*
fin_diccionario

Así mismo, también queda pendiente el aprovechamiento de las relaciones entre los términos del tesauro que vienen dadas en la propia estructura de cada término. De esta manera se incluiría información semántica en el diccionario.
Se puede ampliar la estructura de palabras, agregando información de cuales son los descriptores (o no descriptores) relacionados en donde aparece; se propone guardar el código numérico que precede al término y que permite identificarlo. Por lo tanto, teniendo una lista asociada a cada palabra con estos códigos, uno puede conocer cuales son los términos a los cuales pertenece tal palabra.

Otra observación a destacar, es el hecho de que no se utilizó sólo un área temática específica a la hora de la elección de los términos de la muestra. Esto, a nuestro modo de ver refuerza aun más la bondad de los resultados obtenidos.

En otro orden, es posible de solucionar algunos de los errores que fueron detectados. Es posible tener una lista en donde aparezcan las siglas que maneja el tesauro (en muchos casos, los tesauros vienen acompañados de este tipo de listas), así como también listas conteniendo los nombres propios.

Como ya se comentó en la Sección 6.2 (Resultados obtenidos), existe un conjunto de términos que comienzan con una palabra cuya categoría gramatical es Adjetivo[4]. Haciendo un análisis de las características de este tipo de adjetivos, se vio que estos pertenecen al conjunto de los adjetivos numerales (PRIMER, DOS, SEGUNDA, etc.).
Como la cantidad de palabras en estas condiciones no son muchas es posible construir una lista auxiliar que las contenga. Se trata de palabras que tanto en masculino como en femenino, aportan un valor cuantificador al sustantivo y que por lo tanto pueden estar por delante del mismo.
Una vez conformada dicha lista es factible de hacer un segundo chequeo con la misma a los efectos de corregir errores.
En definitiva, lo que se está proponiendo es ampliar el lexicón de base.

También se puede considerar términos que tienen la estructura

$$\text{PALABRA}_1 \quad \text{PREPOSICION} \quad \text{DETERMINANTE} \quad \text{PALABRA}_2 \quad \text{PALABRA}_3$$

De la manera en que se implementó, a la hora de elegir las funciones necesarias para el cálculo del género y número, como existe un determinante, para inferir el género y número de PALABRA_2 se emplea la función *gen_num_art*[5], dejando que el análisis de PALABRA_3 se realice en forma independiente del contexto por la función *gen_num_adj*[6].

Se puede ampliar el conjunto de funciones auxiliares construyendo otra función que permita verificar/obtener el género y el número para el adjetivo PALABRA_3 de manera tal que concuerde con el del sustantivo PALABRA_2 al que califica.
En este caso, se propone una nueva función *gen_num_art_adj* que se encargue de chequear dicha concordancia.

Otra posible extensión, es el tratamiento de las **locuciones prepositivas**. Como se recordará, se presentó un ejemplo de término con el contenido

DESPUES DEL 2000

donde existe un problema con la palabra DESPUES. Dicha palabra forma parte de una locución prepositiva y como se mencionó oportunamente, este tipo de términos no fueron tratados en este trabajo y por lo tanto se detecta un error, puesto que toma la palabra DESPUES como un sustantivo.
Como los casos de este tipo de complementos prepositivos están perfectamente determinados, el ampliar este trabajo para su tratamiento no sería demasiado complejo.

Otro tema es el tratamiento de los verbos en infinitivo. No se ha encontrado una manera eficiente de evitar la construcción de su forma plural. Estos son considerados como sustantivo-masculino-singular y al querer construir su forma plural se genera un error.
El problema es que no se puede tomar una determinación genérica a partir de las terminaciones, puesto que existen palabras con terminación _ER, _AR como LIDER , HOGAR que sí son palabras con los rasgos antes mencionados y por lo tanto tiene sentido construir su forma plural; mientras que por otro lado se tienen palabras tales como APRENDER y ESCRIBIR, donde se trata de verbos y por lo tanto no ocurre lo mismo.
Hay por otro lado algunos casos en que para un verbo en infinitivo tiene sentido hallar su plural como MILITAR, por ejemplo, donde construir su plural no esta mal (MILITARES). Aunque se está nuevamente el caso de palabras que tiene dos significados diferentes según su contexto.

Entonces, otra posible extensión del trabajo es, además de dar un peso a las reglas que ayudan a deducir ciertos rasgos, asociar a cada palabra una cierta ***probabilidad*** a la asignación de rasgos a palabras. De este modo se podría solucionar el reconocimiento de palabras terminadas en _AR , _ER, _IR, que tienen una mayor ***probabilidad*** de ser verbos en infinitivo

[4] pueden ser subtérminos

[5] función que calcula el género y el número de un sustantivo en base al determiante que lo precede

[6] función que calcula el género y el número de un adjetivo en base a los morfemas derivativos

que sustantivos. A su vez, sirve también para tentar una solución al problema de los adjetivos numerales comentados anteriormente.

De esta manera, deben existir criterios para que dependiendo de éstos valores de probabilidad, se calcule o no la forma plural, etc.

BIBLIOGRAFÍA

[ACC95] Accuosto, Pablo
Selección de descriptores de un tesauro a partir de una consulta temática en lenguaje natural. Reporte interno, Facultad de Ingeniería 1995.

[BEL84] Bello, Andrés.
Gramática de la lengua castellana. Colección Edaf Universitaria 1988.

[BOS91] Bosque, Ignacio
Las categorías gramaticales. Relaciones y diferencias. Editorial Síntesis 1991.

[CAL83] Calzolari, Nicoletta
The dictionary and the thesaurus can be combined. Publicación 1983.

[COC91] Coch, J. ; Charles, V. ; David, R. ; Monze, G.
Biblix : génération rapide d'interfaces documentaires en langage naturel.
Documento producido por las empresas francesas GSI-Erli y Cybernetix. Marsella 1991.

[GUI67] Guiraud, Pierre
La Gramática. Editorial Eudeba - Segunda Edición 1967.

[MAD89] Madrazo, P.G. ; Moragón, C.
Aprende tu sólo Gramática. Ediciones Pirámide, Madrid 1989

[McE92] Mc Enery, A.M.
Computational Linguistics. A handbook & toolbox for Natural Language Proccesing, 1992.

[OGO93] Ogonowski, A. ; Herviou, M.L. ; Dauphin, E.
Tools for extracting and structuring knowledge from texts. Publicación, 1993.

[RAE31] Real Academia Española
Gramática de la Lengua Española - Edición 1931

[RAE86] Real Academia Española
Esbozo de una Nueva Gramática de la Lengua Española - Edición 1986

[SAN95] Sanchez Ladrón de Guevara, A ; García Jumela, F.
Una contribución al procesamiento del plural en español : un algoritmo de singularización - Boletín Nº17 de la SEPLN, 1995.

[SPN84] Tesauro Spines - Un vocabulario controlado y estructurado para el tratamiento de la información sobre ciencia y tecnología para el desarrollo.
Edición española 1984 - Volumen I

[VER94] Vergne, Jaques
A non-recursive sentence segmentation, applied to parsing of a linear complexity in time. Publicación 1994.

[WIE86] Wierzbicka, A
What's a Noun (or How do Nouns differ in meaning from Adjetives) - Studies in Language, 1996.

APÉNDICE

A.1 Algunas pruebas

A continuación se presentan una serie de entradas del Tesauro Spines[7] y el correspondiente análisis para cada palabra.

Prueba 1.-

```
2986 ANALISIS ARMONICO 20
        EN HARMONIC ANALYSIS
        FR ANALYSE HARMINIQUE
        < ANALISIS FUNCIONAL
;

EN CURSO 30
        => PRESENTE
;

7701 FIEBRE AMARILLA 25
        EN YELLOW FEVER
        FR FIEVRE JAUNE
        < ENFERMEDADES PRODUCIDAS POR VIRUS
;

LENGUAJES ARTIFICIALES 18 20
        : ALGORITMOS
;

MAQUINAS DE ESCRIBIR DE CONSOLA 18
        => CONSOLAS DE COMPUTADORES
;
```

El pasado juego de descriptores genera la siguiente salida :

```
ANALISIS esta en Singular y en Plural
ARMONICO esta en Singular y ARMONICOS es su Plural
  Categoria de ARMONICO : ADJETIVO
  Categoria de ANALISIS : SUSTANTIVO
 ARMONICO es de genero Masculino y ARMONICA es su Femenino
 ARMONICO es de genero Masculino
 ANALISIS es de genero Masculino
ANALISIS esta en Singular y en Plural
FUNCIONAL esta en Singular y FUNCIONALES es su Plural
  Categoria de FUNCIONAL : ADJETIVO
  Categoria de ANALISIS : SUSTANTIVO
 FUNCIONAL es de genero Feme / Masc
 ANALISIS es de genero Femenino
Descriptor Valido !!

CURSO esta en Singular y CURSOS es su Plural
 CURSO es de genero Masculino
  Categoria de CURSO : SUSTANTIVO
PRESENTE esta en Singular y PRESENTES es su Plural
 PRESENTE es de genero Masculino
  Categoria de PRESENTE : SUSTANTIVO
No Descriptor Valido !!

FIEBRE esta en Singular y FIEBRES es su Plural
AMARILLA esta en Singular y AMARILLAS es su Plural
  Categoria de AMARILLA : ADJETIVO
```

[7] En algunos casos se eliminaron ciertos términos en la entrada original para no hacer pesado el seguimiento de la ejecución, puesto que al no tener almacenada la palabra reconocida y su información, ésta es analizada tantas veces como aparezca en el término. Cabe además notar que en muchos casos se comprueba que una misma palabra tiene asociado dos géneros (lo que es resuelto con los pesos).

Categoria de FIEBRE : SUSTANTIVO
AMARILLA es de genero Femenino y AMARILLO es su Masculino
AMARILLA es de genero Femenino
FIEBRE es de genero Femenino
VIRUS esta en Singular y en Plural
VIRUS es de genero Masculino
Categoria de VIRUS : SUSTANTIVO
ENFERMEDADES esta en Plural y ENFERMEDAD es su Singular
PRODUCIDAS esta en Plural y PRODUCIDA es su Singular
Categoria de PRODUCIDA : ADJETIVO
Categoria de ENFERMEDAD : SUSTANTIVO
PRODUCIDA es de genero Femenino y PRODUCIDO es su Masculino
PRODUCIDA es de genero Femenino
ENFERMEDAD es de genero Femenino
Descriptor Valido !!

LENGUAJES esta en Plural y LENGUAJE es su Singular
ARTIFICIALES esta en Plural y ARTIFICIAL es su Singular
Categoria de ARTIFICIAL : ADJETIVO
Categoria de LENGUAJE : SUSTANTIVO
ARTIFICIAL es de genero Feme / Masc
LENGUAJE es de genero Masculino
ALGORITMOS esta en Plural y ALGORITMO es su Singular
ALGORITMO es de genero Masculino
Categoria de ALGORITMO : SUSTANTIVO
No Descriptor Valido !!

ESCRIBIR esta en Singular y ESCRIBIRES es su Plural
ESCRIBIR es de genero Masculino
Categoria de ESCRIBIR : SUSTANTIVO
CONSOLA esta en Singular y CONSOLAS es su Plural
CONSOLA es de genero Femenino
Categoria de CONSOLA : SUSTANTIVO
MAQUINAS esta en Plural y MAQUINA es su Singular
MAQUINA es de genero Femenino
Categoria de MAQUINA : SUSTANTIVO
COMPUTADORES esta en Plural y COMPUTADOR es su Singular
COMPUTADOR es de genero Masculino
Categoria de COMPUTADOR : SUSTANTIVO
CONSOLAS esta en Plural y CONSOLA es su Singular
CONSOLA es de genero Femenino
Categoria de CONSOLA : SUSTANTIVO
No Descriptor Valido !!

Cantidad de Descriptores analizados : 2
Cantidad de NO Descriptores analizados : 3
Cantidad de Palabras analizadas : 19 (contando repetidas)

Prueba 2.-

```
AGENTES QUIMICOS MILITARES 34
        => ARMAS QUIMICAS Y BIOLOGICAS
;

MANI 24
        => CACAHUETE (ARBOL)
;

4063 NAVES ESPACIALES TRIPULADAS 26
        EN MANNED SPACECRAFT
        FR SPATIONEFS HABITES
        < NAVES ESPACIALES
        - VEHICULOS INTERPLANETARIOS TRIPULADOS
        - SATELITES ARTIFICIALES
        ;
```

El pasado juego de descriptores genera la siguiente salida :

```
AGENTES esta en Plural y AGENTE es su Singular
QUIMICOS esta en Plural y QUIMICO es su Singular
MILITARES esta en Plural y MILITAR es su Singular
  Categoria de QUIMICO : ADJETIVO
  Categoria de MILITAR : ADJETIVO
  Categoria de AGENTE : SUSTANTIVO
 QUIMICO es de genero Masculino y QUIMICA es su Femenino
 QUIMICO es de genero Masculino
 MILITAR es de genero Masculino
 AGENTE es de genero Masculino
ARMAS esta en Plural y ARMA es su Singular
QUIMICAS esta en Plural y QUIMICA es su Singular
BIOLOGICAS esta en Plural y BIOLOGICA es su Singular
  Categoria de QUIMICA : ADJETIVO
  Categoria de BIOLOGICA : ADJETIVO
  Categoria de ARMA : SUSTANTIVO
 QUIMICA es de genero Femenino y QUIMICO es su Masculino
 BIOLOGICA es de genero Femenino y BIOLOGICO es su Masculino
 QUIMICA es de genero Femenino
 BIOLOGICA es de genero Femenino
 ARMA es de genero Femenino
No Descriptor Valido !!

MANI esta en Singular y MANIES es su Plural
 MANI es de genero Masculino
  Categoria de MANI : SUSTANTIVO
ARBOL esta en Singular y ARBOLES es su Plural
 ARBOL es de genero Masculino
  Categoria de ARBOL : SUSTANTIVO
CACAHUETE esta en Singular y CACAHUETES es su Plural
 CACAHUETE es de genero Masculino
  Categoria de CACAHUETE : SUSTANTIVO
No Descriptor Valido !!

NAVES esta en Plural y NAVE es su Singular
ESPACIALES esta en Plural y ESPACIAL es su Singular
TRIPULADAS esta en Plural y TRIPULADA es su Singular
  Categoria de ESPACIAL : ADJETIVO
  Categoria de TRIPULADA : ADJETIVO
  Categoria de NAVE : SUSTANTIVO
 TRIPULADA es de genero Femenino y TRIPULADO es su Masculino
 ESPACIAL es de genero Feme / Masc
 TRIPULADA es de genero Femenino
 NAVE es de genero Femenino
NAVES esta en Plural y NAVE es su Singular
ESPACIALES esta en Plural y ESPACIAL es su Singular
  Categoria de ESPACIAL : ADJETIVO
  Categoria de NAVE : SUSTANTIVO
```

ESPACIAL es de genero Feme / Masc
NAVE es de genero Masculino
VEHICULOS esta en Plural y VEHICULO es su Singular
INTERPLANETARIOS esta en Plural y INTERPLANETARIO es su Singular
TRIPULADOS esta en Plural y TRIPULADO es su Singular
Categoria de INTERPLANETARIO : ADJETIVO
Categoria de TRIPULADO : ADJETIVO
Categoria de VEHICULO : SUSTANTIVO
INTERPLANETARIO es de genero Masculino y INTERPLANETARIA es su Femenino
TRIPULADO es de genero Masculino y TRIPULADA es su Femenino
INTERPLANETARIO es de genero Masculino
TRIPULADO es de genero Masculino
VEHICULO es de genero Masculino
SATELITES esta en Plural y SATELITE es su Singular
ARTIFICIALES esta en Plural y ARTIFICIAL es su Singular
Categoria de ARTIFICIAL : ADJETIVO
Categoria de SATELITE : SUSTANTIVO
ARTIFICIAL es de genero Feme / Masc
SATELITE es de genero Masculino
Descriptor Valido !!

Cantidad de Descriptores analizados : 1
Cantidad de NO Descriptores analizados : 2
Cantidad de Palabras analizadas : 19 (contando repetidas)

A.2 Lexicón

```
/*      ========================================================
        Conjunto base - Lexicográfico

        Diccionario que contiene las preposiciones, conjunciones y determinantes

        ========================================================      */

%{
#include "y.tab.h"
#include "auxilio.h"
%}

nro     [0-9]
letra   [A-Z]
caresp  [ \'\&\+]

%%

{nro}{1,4}      { yylval.n = atoi(yytext);     /*      Reconoce números para las Areas temáticas  */
                  return(NRO);
                }

/*      Preposiciones           */

A               { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

ANTE            { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

BAJO            { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

CON             { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

CONTRA          { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

DE              { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

DESDE           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }
```

```
EN              { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

ENTRE           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

HACIA           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

PARA            { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

POR             { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

SEGUN           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

SIN             { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

SOBRE           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

MEDIANTE        { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

DURANTE         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

SALVO           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

INCLUSO         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

EXCEPTO         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PREP);
                }

/*      Preposiciones  contractivas     */

AL              { yylval.st = ((char *)malloc(yyleng+1));
```

```
                  strcpy(yylval.st,yytext);
                  return(PCONTRAC);
                }

DEL               { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(PCONTRAC);
                }

/*      Conjunciones            *7

Y                 { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(Y);
                }

E                 { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(Y);
                }

/*      Adverbio de negación  NO        */

NO                return(NO);


/*      Determinantes           */

EL                { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

LA                { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

LAS               { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

LOS               { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

UN                { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

UNA               { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }
```

```
UNAS        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

UNOS        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESTE        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESTO        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESTOS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESTA        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESTAS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESE         { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESO         { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESOS        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESA         { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

ESAS        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

AQUEL       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

AQUELLO     { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
```

```
                }

AQUELLOS        { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

AQUELLA         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

AQUELLAS        { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

ALGUN           { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

ALGUNO          { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

ALGUNOS         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

ALGUNA          { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

ALGUNAS         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

NINGUN          { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

NINGUNO         { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

NINGUNOS        { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }
```

```
NINGUNA     { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

NINGUNAS    { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

MUCHO       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

MUCHOS      { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

MUCHA       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

MUCHAS      { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

POCO        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

POCOS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

POCA        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

POCAS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

OTRO        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

OTROS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

OTRA        { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
            }

OTRAS       { yylval.st = ((char *)malloc(yyleng+1));
              strcpy(yylval.st,yytext);
              return(DET);
```

```
                    }

DEMASIADO       { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

DEMASIADOS      { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

DEMASIADA       { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }

DEMASIADAS      { yylval.st = ((char *)malloc(yyleng+1));
                  strcpy(yylval.st,yytext);
                  return(DET);
                }



/*          Otras construcciones        */


FR                  return(FR);

I\+D                return(CTID);
C\+T                return(CTID);

[ \t\n]             ;
\;                  return(PYC);
\(                  return(PARABR);
\)                  return(PARCIE);



/*          Símbolos de las relaciones  del Tesauro      */


\<                  return(TG);
\>                  return(TE);
\=                  return(USADOPOR);
\-                  return(TR);
\*                  return(VIENEDE);
\=>                 return(USESE);
\:                  return(VEASE);

\.                  return(PTO);
```

```
/*          Palabras            */

{letra}*            { yylval.st = ((char *)malloc(yyleng+1));
                      strcpy(yylval.st,yytext);
                      return(PALABRA);
                    }

{caresp}            { yylval.st = ((char *)malloc(yyleng+1));
                      strcpy(yylval.st,yytext);
                      return(CAR_ESP);
                    }


({letra}|{nro})*    { yylval.st = ((char *)malloc(yyleng+1));
                      strcpy(yylval.st,yytext);
                      return(PAL_DIG);
                    }


({nro}*\-{nro}*)    { yylval.st = ((char *)malloc(yyleng+1));
                      strcpy(yylval.st,yytext);
                      return(NUM_GUI_NUM);
                    }


({letra}*\-{letra}*) { yylval.st = ((char *)malloc(yyleng+1));
                      strcpy(yylval.st,yytext);
                      return(PAL_GUI_PAL);
                    }


%%
```