4 Desarrollo de datos lingüísticos

4.1 Introducción

En este capítulo se describe el proceso que se siguió al seleccionar los recursos que permitieron implementar el sistema para poder evaluar el modelo desarrollado. Primero, se describen las razones para seleccionar el corpus más adecuado y disponible para correr las pruebas. La decisión se tomó después del análisis, apoyado en programas específicos de verificación y el análisis visual y manual necesario. En segundo lugar, ante la necesidad de probarlo con texto libre, se comenta la inclusión: a) del etiquetador TnT, al cual se le entrenó para el Español; b) del diccionario de sinónimos del Laboratorio de Lenguaje Natural del CIC-IPN que fue necesario corregir y validar manualmente; c) del diccionario de escenarios construido obteniendo la información de relaciones del diccionario semántico EuroWordNet en Español. En el capítulo siguiente se presentarán los resultados obtenidos.

4.2 Selección del corpus a utilizar

Para la obtención del corpus lingüístico se tenían inicialmente las siguientes alternativas:

a) Lecturas de Libros de Texto Gratuitos en Español obtenidos de la página Web de la Secretaría de Educación Pública en México. Se esperaba como ventajas la seguridad (relativa) de ser textos bien redactados (sin faltas de ortografía) y revisados, con buen nivel de lenguaje; se tendría que verificar el permiso para utilizarlo en el trabajo de investigación (Copyright). Después de un análisis de los textos se observó: una redacción y vocabulario demasiado elemental; NULA presencia del fenómeno de anáfora hasta el cuarto año de primaria; a partir del quinto año de primaria uso de anáfora directa por medio de pronombres (muy elemental); casi nula presencia del fenómeno de anáfora indirecta (menor al 1 %).

Se descartó, de acuerdo a lo anterior, por no ser representativo del fenómeno a analizar.

b) LexEsp (Léxico informatizado del español) – es un corpus en Español etiquetado con alrededor de 5 millones de palabras en texto libre, compilado por: el Departamento de Psicología de la Universidad de Oviedo, el Grupo Lingüístico de la Universidad de Barcelona y el Grupo para el tratamiento del Lenguaje de la Universidad Politécnica de Cataluña, recogido entre los años 1978 y 1995. Tiene como ventajas el poder reportar resultados con un corpus que ha sido utilizado por otros investigadores [Muñoz, 2000]; al ser etiquetado facilita el desarrollo del prototipo inicial; otra ventaja es que se tiene el permiso para utilizarlo en el trabajo de investigación (Copyright). Se tomó una muestra al azar representativa de 15 archivos y desarrollaron programas para efectuar la revisión observando: errores de captura del texto (palabras mal escritas, puntuación, etc.); el etiquetado fue realizado con un etiquetador automático con errores ante palabras ambiguas; incongruencia entre el manual y los archivos en las claves de etiquetado; y la presencia de hasta 14 temas diferentes en cada archivo sin separación marcada. Se descartó por los errores mencionados y además porque para el proyecto es deseable que cada archivo contenga un solo artículo o tema debido a la importancia del contexto lingüístico.

c) Corpus en español que se ha ido compilando en el Laboratorio de Lenguaje Natural del CIC-IPN de noticias en periódicos publicadas en el Web, en texto libre. Tiene como ventaja el poder reportar resultados con un corpus que ha sido utilizado por otros investigadores [Galicia-Haro et al., 1999]. Su desventaja es que no es un texto libre de errores (ortográficos y de redacción); la presencia de temas diferentes en cada archivo sin separación marcada; y no está etiquetado. Se descartó por los errores mencionados y además porque para el proyecto es deseable que cada archivo contenga un solo artículo o tema debido a la importancia del contexto lingüístico.

Ante la situación mostrada se localizó un corpus etiquetado “Corpus CLiC-TALP que se planea esté formado por 1,000,000 de palabras etiquetadas, desambiguadas y validadas

manualmente", con fecha de la ultima actualización en Enero del 2002. Los propietarios del corpus son la Universidad Politécnica de Cataluña y la Universidad de Barcelona en España que ofrecen la cesión gratuita de derechos a investigadores. Se contactó a los responsables y obtuvo una muestra representativa del corpus validada manualmente (el proyecto está en proceso).

El corpus CLiC–TALP proviene de dos fuentes diferentes. Por una parte recoge una muestra representativa (de 500.000 palabras) de un corpus de prensa de 7 millones de palabras cedido por el periódico *La Vanguardia*. Por otra, recoge una muestra (también de 500.000 palabras) del corpus LexEsp, que es un corpus de 5 millones de palabras, representativo del español estándar escrito porque presenta varios estilos narrativos, procedentes de distintas fuentes (literatura, prensa, etc.) e incluye también muestras tanto del español peninsular como del de América. Recoge un número reducido de palabras por obra y no más de tres obras por autor. Las fuentes son las que aparecen en la tabla 6.

Fuentes	**Porcentaje**
Narrativa	40
Divulgación científica	10
Ensayo	10
Prensa	25
Semanarios	10
Prensa deportiva	5

Tabla 6 Fuentes de LexEsp

Recoge muestras de 329 novelas con unas 6.000 palabras por obra, aproximadamente. Las revistas de divulgación científica utilizadas han sido *Muy interesante, Mundo científico* e *Investigación y ciencia*, así como algunos artículos de Divulgación publicados en suplementos de periódicos como *El País* y *ABC*. Los fragmentos de ensayo provienen de unas 88 obras, a razón de unas 5.700 palabras por obra. La parte procedente de prensa se ha obtenido de *El País, ABC, El Mundo, El Periódico, Diario 16, El Independiente* y *La Vanguardia*. Hay que reseñar que esta parte se compone de otras tres: editoriales (15%), articulistas (50%) y noticias (35%). Los semanarios utilizados han sido *Cambio 16, Interviú, Época* y *Tiempo*. Por último la parte de la prensa deportiva proviene de las publicaciones *As, Marca* y *Mundo Deportivo.* La parte del corpus CLiC-TALP extraída de LexEsp aparece en diferentes archivos cuyos nombres contienen

una letra inicial seguida de un número. La letra se corresponde con las distintas fuentes utilizadas, de modo que es posible conocer el tipo de texto. En la tabla 7 se presenta la relación entre el nombre del archivo y su contenido.

Letra Inicial	Contenido
a	Articulistas
e	Ensayo
d	prensa deportiva
dc	Divulgación científica
c	suplementos de ciencia
ed	Editoriales
n	Noticias
r	Semanarios
t	Narrativa

Tabla 7 Contenido de la parte de LexEsP

Se desarrollaron programas de verificación y después de analizarlo se encontraron errores de captura (menor al 2%) y errores de documentación; estos errores se notificaron a Montserrat Civit (responsable y contacto) quien corrigió el manual agradeciendo las observaciones sugeridas. Se observó que en cada archivo había cuando mucho tres temas, pudiendo dividirse manualmente hasta obtener un solo artículo o tema. Por lo anterior se consideró más adecuado como alternativa, además de ofrecer la ventaja de evitar el preprocesamiento para obtener la información morfosintáctica de las expresiones lingüísticas para la prueba del prototipo inicial.

4.3 Adecuación y entrenamiento del etiquetador TnT

El etiquetador de partes de la oración es un módulo requerido como etapa de preprocesamiento cuando se necesita trabajar con texto libre, aunque no fue necesario para el prototipo inicial porque se trabajo directamente con el corpus etiquetado CLiC-TALP. El etiquetador es un programa que acepta texto libre o no preparado como entrada y añade a cada unidad léxica (token) una etiqueta que especifica sus propiedades gramaticales como categoría, número, persona, etc. Para este trabajo se localizó y obtuvo uno de los mejores y más populares —TnT tagger— que había sido probado exhaustivamente para el Inglés y Alemán pero no para el Español.

El etiquetador TnT (TnT es el acrónimo de Trigrams'n'Tags) es de tipo estadístico que puede ser entrenado para diferentes lenguajes y conjuntos de etiquetas dependiendo de un corpus etiquetado. Su exactitud promedio está alrededor del 96.5% para palabras conocidas y del 87% para palabras desconocidas. Fue necesario realizar una evaluación con el fin de conocer que tan adecuado es para el Español y cuanto afectaría al rendimiento global del sistema obteniendo un rendimiento del 96.5%, igual al promedio, para palabras conocidas y del 79.8% para palabras desconocidas [Morales y Gelbukh, 2003].

La diferencia del 16.7% entre la precisión obtenida con palabras conocidas y desconocidas hizo necesario analizar la causa de los errores de etiquetado en el Español encontrando tres causas principales:

(**I**) El orden más libre de palabras en **la posición del adjetivo** por ejemplo:

(84) Juan compró un **nuevo** carro (otro carro, tal vez usado; posición poco usada)

(85) Juan compró un carro **nuevo** (un carro recién fabricado; posición más usada)

A continuación se presentan ejemplos de la corrida de TnT (resaltando la ocurrencia del error con **negrita**).

Ejemplo 1	
Etiquetado TnT	**Debe ser**
los DA0MP0	los DA0MP0
sórdidos **NCMP000**	sórdidos AQ0MP0
arenales AQ0CP0	arenales **NCMP000**

Ejemplo 2	
Etiquetado TnT	**Debe ser**
precarios **NCMP000**	precarios AQ0MP0
tenderetes AQ0CP0	tenderetes **NCMP000**
de SPS00	de SPS00
cartón NCMS000	cartón NCMS000

(**II**) **Ambigüedad por homografía** (formas iguales de palabras) con diferentes significados o funciones gramaticales, por ejemplo:

(86) Yo **bajo**$_1$ con el hombre **bajo**$_2$ a tocar el **bajo**$_3$ **bajo**$_4$ las escaleras (**bajo**$_1$ verbo, **bajo**$_2$ adjetivo, **bajo**$_3$ nombre de instrumento musical y **bajo**$_4$ adverbio de lugar)

(87) Deja el **sobre**$_1$ que **sobre**$_2$ **sobre**$_3$ la mesa (**sobre**$_1$ nombre, **sobre**$_2$ verbo y **sobre**$_3$ adverbio de lugar)

A continuación se presentan dos ejemplos de la corrida de TnT (resaltando la ocurrencia del error con **negrita)**.

Ejemplo 3	
Etiquetado TnT	**Debe ser**
y CC	y CC
recuerdo NCMS000	**recuerdo VMIP1S0**
aquel DD0MS0	aquel DD0MS0
almuerzo NCMS000	almuerzo NCMS000
conmovedor AQ0MS0	conmovedor AQ0MS0

En el ejemplo 3, la forma de palabra **recuerdo** puede ser el verbo recordar en tiempo pasado, primera persona del singular o un nombre que denota la memoria de un hecho del pasado.

Ejemplo 4	
Etiquetado TnT	**Debe ser**
se P0300000	se P0300000
sujeta VMIP3S0	sujeta VMIP3S0
la DA0FS0	la DA0FS0
cola NCFS000	cola NCFS000
con SPS00	con SPS00
la DA0FS0	la DA0FS0
boca NCFS000	boca NCFS000
y CC	y CC
rueda NCFS000	**rueda VMIP3S0**
como CS	como CS
una DI0FS0	una DI0FS0
rueda NCFS000	*rueda* NCFS000

En el ejemplo 4, hay dos ocurrencias de la forma de palabra **rueda** en la misma oración; en la primera ocurrencia es el verbo rodar en tiempo presente, tercera persona del singular, mal identificado por TnT; en la segunda ocurrencia es un nombre bien identificado por TnT.

(**III**) **Afijos o terminaciones iguales** de las palabras con diferente función

Ejemplo 5	
Etiquetado TnT	**Debe ser**
Comimos VMIS1P0	Comimos VMIS1P0
un DI0MS0	un DI0MS0
arroz AQ0CS0	**arroz NCMS000**
con SPS00	con SPS00
pollo NCMS000	pollo NCMS000

El error de TnT se debe a que el afijo –oz es común para nombres (arroz, voz) y adjetivos (atroz, feroz, portavoz).

Ejemplo 6	
Etiquetado TnT	**Debe ser**
la DA0FS0	la DA0FS0
nariz NCFS000	nariz NCFS000
en SPS00	en SPS00
forma NCFS000	forma NCFS000
de SPS00	de SPS00
taco NCMS000	taco NCMS000
de SPS00	de SPS00
billar **VMN0000**	billar **NCMS000**

El error de TnT se debe a que el afijo –ar denota el infinitivo de verbos de la primera conjugación pero también es común en nombres (billar, azúcar, telar) y adjetivos (espectacular, lumbar, estándar, molecular).

En el ejemplo 7 el afijo –ía se utiliza para el tiempo pasado simple del modo indicativo en la segunda conjugación para la primera y tercera persona del singular. En el ejemplo 8 el afijo –aba se utiliza para la primera y tercera persona del singular en el modo indicativo y tiempo pasado imperfecto de la primera conjugación. El error de TnT es en la identificación de la persona (intercambio de primera y tercera persona). Estos ejemplos muestran la clase de ambigüedad más común manifestada en la conjugación de verbos en Español; el problema aumenta con el uso de cortesía o político del "ustedes" en lugar de "vosotros" cada vez más

común en el Español moderno porque los afijos son iguales para la segunda y tercera persona del plural, como se puede observar en la tabla 8 donde se han concentrado los casos de ambigüedad.

Ejemplo 7	
Etiquetado TnT	**Debe ser**
Pero CC	Pero CC
la DA0FS0	*la DA0FS0*
existencia NCFS000	*existencia NCFS000*
de SPS00	de SPS00
dos DN0CP0	dos DN0CP0
recién RG	recién RG
nacidos AQ0MPP	nacidos AQ0MPP
en SPS00	en SPS00
la DA0FS0	la DA0FS0
misma DI0FS0	misma DI0FS0
caja NCFS000	caja NCFS000
sólo RG	sólo RG
pod**ía VMII1S0**	pod**ía VMII3S0**
deberse VMN0000	deberse VMN0000
a SPS00	a SPS00
un DI0MS0	un DI0MS0
descuido NCMS000	descuido NCMS000
de SPS00	de SPS00
fábrica NCFS000	fábrica NCFS000

Ejemplo 8	
Etiquetado TnT	**Debe ser**
Me PP**1**CS000	*Me* PP**1**CS000
obsesion**aba VMII3S0**	obsesion**aba VMII1S0**
la DA0FS0	la DA0FS0
imagen NCFS000	imagen NCFS000
del SPCMS	del SPCMS
pobre AQ0CS0	pobre AQ0CS0
Niño_Dios NP00000	Niño_Dios NP00000
rechazado AQ0MSP	rechazado AQ0MSP

				Afijos de los modelos de Conjugación		
Modo	**Tiempo**	**Persona**	**Número**	**1st**	**2nd**	**3rd**
Indicativo	presente	2	plural	-an	-en	-en
		3	plural	-an	-en	-en
	pasado simple	1	singular	-aba	-ía	-ía
		3	singular	-aba	-ía	-ía
		2	plural	-ban	-ían	-ían
		3	plural	-ban	-ían	-ían
	pasado indefinido	2	plural	-ron	-ieron	-ieron
		3	plural	-ron	-ieron	-ieron
	futuro simple	2	singular	-rán	-erán	-irán
		3	singular	-rán	-éran	-irán
	condicional simple	1	singular	-ría	-ería	-iría
		2	singular	-ría	-ería	-iría
		2	plural	-rían	-erían	-irían
		3	plural	-rían	-erían	-irían
Subjuntivo	presente	1	singular	-e	-a	-a
		3	singular	-e	-a	-a
		2	plural	-en	-an	-an
		3	plural	-en	-an	-an
	pasado simple	1	singular	-ara	-iera	-iera
		3	singular	-ara	-iera	-iera
		2	plural	-aran	-ieran	-ieran
		3	plural	-aran	-ieran	-ieran
	futuro simple	1	singular	-are	-iere	-iere
		3	singular	-are	-iere	-iere
		2	plural	-aren	-ieren	-ieren
		3	plural	-aren	-ieren	-ieren

Tabla 8 Ambigüedad en la conjugación de verbos

El error tipo (I) por posición del adjetivo, ver ejemplo (84) y (85), es poco usual por lo que se considera que no repercutirá significativamente en los resultados a obtener cuando se integre al sistema; el error tipo (II), ejemplos (86) y (87), es común a otras lenguas (Inglés por lo menos) y dependerá del entrenamiento ("casos conocidos") para que "aprenda" a discriminarlos; el error tipo(III), ejemplos de TnT 5 y 6, necesariamente depende del entrenamiento por lo que es necesario aumentar el vocabulario del etiquetador, número de archivos de entrenamiento, para solucionarlo; finalmente, el error tipo (III), ejemplos de TnT 7 y 8, no puede ser solucionado por el etiquetador en la situación actual y se requiere de una etapa posterior de análisis apoyado en la

concordancia con el sujeto de la oración para etiquetar correctamente el texto [Morales y Gelbukh, 2003].

Como se ha observado, los errores de etiquetado repercuten directamente la precisión del método con archivos de texto libre, lo que abre un área de oportunidad para mejorar el preprocesamiento de etiquetado y como consecuencia el método de resolución de la anáfora indirecta. Mención especial requiere el error tipo (III), de los ejemplos de TnT 7 y 8, que limita el uso del etiquetador para la resolución de la anáfora directa donde es indispensable la concordancia del sujeto con el verbo; sin embargo, no afecta al método de resolución de la anáfora indirecta desarrollado porque sólo utiliza la categoría gramatical.

4.4 Preparación del diccionario de sinónimos

El diccionario de sinónimos inicial se recibió del Laboratorio de Lenguaje Natural del CIC-IPN; este diccionario había sido obtenido con un escáner a partir del diccionario de sinónimos Océano. La revisión que se hizo mostró que estaba pendiente la verificación del proceso y corrección de errores. Ejemplos de los errores más comunes se presentan en la tabla 9 donde se observa que algunos errores habían quedado marcados con la letra **q** (sinónimos en fila 1, 2 y 4) y otros sólo podrían ser detectados por verificación visual directa (entrada en fila 3). Se puede observar también que hay una entrada para el singular y otra para el plural (ÉLITE y ÉLITES en filas 1 y 2); que algunos sinónimos están formados por una expresión idiomática (BUEN**AqS**OCIEDAD en filas 1 y 2); que existen errores de detección (ÉMU**I**O en lugar de ÉMU**L**O en fila 3; PR**EqD**ILECTO en lugar de PR**ED**ILECTO; **qC**ONO en lugar de **ÍC**ONO en fila 4); que el diccionario utiliza mayúsculas mientras que el texto libre "normalmente" utiliza minúsculas y sólo en la primera palabra de una oración, nombres propios, abreviaturas o siglas utiliza mayúsculas.

Se hizo la verificación y corrección de errores, apoyado con programas y manualmente, para lograr un diccionario que: utilice minúsculas; considere la limitante tamaño de registro del sistema de archivos de texto (en computadoras personales de 1022 caracteres) y en lo posible mantenga una sola entrada para nombres, masculino singular, e infinitivo para verbos; no contenga expresiones idiomáticas porque las comparaciones se hacen con una palabra núcleo de

la expresión nominal. Se alcanzó la corrección de los errores en la tabla 9 obteniendo los resultados mostrados en la tabla 10.

La corrección redujo el número de entradas, de palabras y de palabras / entrada en el diccionario; además, si se considera "idealmente" la consulta al diccionario de una palabra que se encuentre a la mitad tendremos también una reducción en el número de accesos lógicos, al archivo en disco, al reducir el número de entradas (17% aprox.), como se muestra (ver 15816 en la fila **Corregido** de la tabla 11).

N°	ENTRADA	SINÓNIMOS
1	ÉLITE	ÉLITES BUENAqSOCIEDAD CÍRCULO CÍRCULOS CREMA CREMAS
2	ÉLITES	ÉLITE BUENAqSOCIEDAD CÍRCULO CÍRCULOS CREMA CREMAS DISTINCIÓN DISTINCIONES ELEGANCIA LAqFLOR LOqMEJOR MUNDANERÍA SELECCIÓN SELECCIONES
3	ÉMUIO	ADVERSARIA ADVERSARIO ADVERSARIOS ANTAGONISTA ANTAGONISTAS COMBATIENTE
4	PREFERIDO	PREqDILECTO ELEGIDO PRIVILEGIADO FAVORITO PRIVADO PROTEGIDO qCONO

Tabla 9 Errores en diccionario de sinónimos

N°	ENTRADA	SINÓNIMOS
1 y 2	élite	círculo crema distinción elegancia selección
3	émulo	adversario antagonista competidor contrario rival opuesto contendiente contrincante
4	preferido	predilecto elegido privilegiado favorito privado protegido ícono

Tabla 10 Corrección de diccionario de sinónimos

	entradas	**palabras**	**accesos**	**Pal/Ent**
Inicial	38245	260986	19123	6.82
Corregido	31632	223028	15816	6.76
Final	31632	223028	586	6.76

Tabla 11 Modificaciones al diccionario de sinónimos

letra	entrada	palabra	acceso	Pal/Ent
a	**3858**	27575	1929	7.15
b	849	6182	425	7.28
c	3857	**27577**	1929	7.15
d	3163	21571	1582	6.82
e	3274	22833	1637	6.97
f	1036	7673	518	7.41
g	750	5609	375	**7.48**
h	662	4848	331	7.32
i	2192	14851	1096	6.78
j	270	2014	135	7.46
k	**6**	**31**	3	5.17
l	787	5806	394	7.38
m	1524	10858	762	7.12
n	346	2427	173	7.01
ñ	14	96	7	6.86
o	563	3972	282	7.06
p	2638	18519	1319	7.02
q	126	827	63	6.56
r	1863	13030	932	6.99
s	1544	10546	772	6.83
t	1195	8409	598	7.04
u	144	1015	72	7.05
v	697	4904	349	7.04
w	11	45	6	4.09
x	18	60	9	**3.33**
y	49	342	25	6.98
z	196	1408	98	7.18
Suma	31632	223028		
Mínimo	6	31	**3**	3.33
Máximo	3858	27577	**1929**	7.48
Promedio	1172	8260	**586**	6.76

Tabla 12 Análisis del diccionario de sinónimos

Sin embargo, el problema del acceso secuencial de archivos continúa presente, lo que provoca lentitud de procesamiento del programa. Para reducir el tiempo de acceso en la búsqueda secuencial se dividió el archivo de sinónimos en varios archivos, de acuerdo a la letra inicial de la palabra de entrada (incluyendo las acentuadas en caso de iniciar con vocal), obteniendo los valores mostrados en la tabla 12. Este análisis permitió visualizar una reducción del tiempo de

acceso al manejar de esta forma el diccionario porque: en el **peor** de los casos, de nombres que inician con la letra "a", se tienen 1989 accesos lógicos contra 15816 sin la división del diccionario logrando reducir casi 8 veces el número de accesos (15816/1989 = 7.95; ver la fila **Máximo** de la tabla 12); en el **mejor** de los casos, letra "k", con sólo 3 accesos lógicos al disco (de acuerdo al tamaño de "buffer" y memoria caché de los discos actuales, con sólo un acceso físico) se obtiene la información requerida (ver la fila **Mínimo** de la tabla 12); y el caso **promedio** con 585 accesos lógicos (ver la fila **Final** de la tabla 11 y la fila **Promedio** de la tabla 12). Con esta forma de manejar el diccionario se logro reducir el tiempo de procesamiento de más de 8 minutos por archivo a menos de 2 minutos por archivo.

4.5 Preparación del diccionario de escenarios

Para la preparación del diccionario de escenarios se tomó como base la información contenida en el diccionario semántico EuroWordNet LE2-4003 WP6.5. Esta versión del diccionario semántico, propiedad del Laboratorio de Lenguaje Natural del CIC-IPN, no se suministra con bibliotecas de funciones para acceder directamente la información (sólo se suministran para WordNet en Inglés en la versión 1.5). Para utilizarlo en Español fue necesario utilizar el archivo de exportación (Export file) en formato de texto ASCII; en la tabla 13 se muestra como ejemplo una entrada del archivo.

Primero se analizó la documentación y se seleccionaron las relaciones necesarias para relacionar núcleos de expresiones nominales (nombres); después se desarrollaron programas de acceso y extracción. El proceso de extracción se validó con un programa de prueba para contar las relaciones, permitiendo probar las rutinas de acceso, y se obtuvieron los resultados mostrados en la tabla 14; en esta tabla se hace la comparación contra las cantidades en la documentación de EuroWordNet observando las diferencias mostradas en la tercera columna.

Las diferencias observadas hicieron necesario verificar la posibilidad de error en los programas, rutinas o en la información del diccionario de EuroWordNet. El análisis condujo a encontrar diferentes tipos de errores en el diccionario que provocan estas diferencias (y corroboran que los programas y rutinas están trabajando bien); además, permitió observar errores adicionales que afectarán, de una forma u otra, la información obtenida para la resolución de la

anáfora indirecta; ejemplos de estos errores enumerados en la primera columna, resaltados con **negrita** en ENTRADA y RELACIONES, se muestran en la tabla 15.

```
0 @5106@ WORD_MEANING
 1 PART_OF_SPEECH "n"
 1 VARIANTS
  2 LITERAL "órgano"
   3 SENSE 4
  2 LITERAL "órgano_vegetal"
   3 SENSE 1
 1 INTERNAL_LINKS
  2 RELATION "has_hyperonym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "cosa"
     5 SENSE 1
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "retoño"
     5 SENSE 2
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "follaje"
     5 SENSE 2
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "cabillo"
     5 SENSE 1
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "offset"
     5 SENSE 4
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "raíz"
     5 SENSE 2
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "lámina"
     5 SENSE 2
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "estructura_reproductiva"
     5 SENSE 1
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "ascocarpo"
     5 SENSE 1
  2 RELATION "has_hyponym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "esporocarpo"
     5 SENSE 1
 1 EQ_LINKS
  2 EQ_RELATION "eq_synonym"
   3 TARGET_ILI
    4 PART_OF_SPEECH "n"
    4 WORDNET_OFFSET 7977350
  2 EQ_RELATION "eq_has_hyperonym"
   3 TARGET_ILI
    4 PART_OF_SPEECH "n"
    4 WORDNET_OFFSET 7976849
```

Tabla 13 Ejemplo de formato de WordNet en Español

RELACIÓN	Documentado	Contado	Diferencia
has_holo_madeof	110	108	2
has_holo_member	427	426	1
has_holo_part	1929	1923	6
has_hyperonym	24608	24507	101
has_hyponym	24608	24507	101
has_mero_madeof	110	108	2
has_mero_member	427	426	1
has_mero_part	1929	1923	6
TOTAL	**54148**	**53926**	**220**

Tabla 14 Relaciones obtenidas de WordNet en Español

N°	ENTRADA	RELACIONES
1	añil	has_hyperonym añil
2	añil	has_hyperonym color
3	añil	has_hyperonym color
4	día_de_la_independencia	has_holo_part julio
5	día_de_la_bandera	has_holo_part junio
6	nuclio	has_hyperonym palabra
7	lugar	has_hyponym P
8	P	has_hyperonym lugar
9	Méjico	has_mero_member mejicano

Tabla 15 Ejemplo de errores de WordNet en Español

Un tipo de error **1** es encontrar alguna relación incorrecta, otro tipo de errores, **2** y **3**, es encontrar entradas duplicadas; en ambos casos el proceso rechaza como incorrecta la relación duplicada, durante la creación del diccionario de escenarios; estos tres tipos de errores son la causa de las diferencias (reducción en el número de relaciones en un 0.4%).

Los errores **4** y **5** son errores de contexto porque para el Inglés de Estados Unidos de Norteamérica ambas entradas están relacionadas con los meses de **julio** y **junio** pero para cualquier otro país la relación es falsa; por ejemplo para México deberían ser septiembre y febrero respectivamente (16 de septiembre y 24 de febrero).

El error tipo **6** representa errores de captura en la entrada provocando que esta relación se pierda porque en el diccionario existe **núcleo** que es lo correcto y no **nuclio** (mal escrito y sin acento); en otras palabras, el programa que busque las relaciones posibles para la entrada núcleo encontrará la entrada relacionada con "*órgano célula cromosoma cromatina centro átomo conjunto importancia mecanismo disco sumista cognición*" pero desconectada de la relación de hiponimia con "*palabra*" debido a este error de captura. Un caso diferente existe con las entrada **médula** y **medula**; ambas acepciones son permitidas por el Diccionario de la Real Academia Española y los desarrolladores de EuroWordNet trataron acertadamente de mantener las variantes registradas pero no lo lograron porque falta esta duplicidad en las entradas sistema_nervioso_central y estructura_neurológica.

Los errores tipo **7** y **8** representan errores de adecuación al idioma porque esta relación no existe (al menos en México) ya que "lugar has_hyponym P" proviene del Inglés "P = Parking" y tiene su equivalente en Español con "E = estacionamiento".

El error tipo **9** se comete por no observar que el diccionario de la Real Academia Española admite la acepciones con "x" y con "j", México y Méjico respectivamente, por lo tanto, se deberían registrar y mantener ambas variantes pero no sucede así; México y todas sus variantes: mexicano, mexiquense, etc. no existen en el diccionario EuroWordNet. Un texto libre donde se registre la acepción con "x" fallará irremediablemente.

El análisis de los errores anteriores alerta sobre: el impacto en la precisión del sistema desarrollado; la necesidad de desarrollar recursos propios para el Español más confiables; y la necesaria participación de México en el desarrollo de sistemas que se promueven como el estándar "de facto" en el mundo.

El siguiente paso, para obtener el diccionario de escenarios, fue la extracción de las relaciones seleccionadas; se muestran los totales en la columna **Relaciones** de la tabla 18 y un ejemplo en la tabla 16. La entrada órgano tomada como ejemplo, muestra cuatro diferentes sentidos de la palabra de acuerdo al contexto utilizado: periodístico, botánico, musical, etc. Habiendo observado que el sistema de resolución de la anáfora indirecta requiere sólo "saber" que existe una relación para establecer el enlace se decidió reducir el tamaño de este diccionario con una estructura y forma de manejo similar a la del diccionario de sinónimos; lo anterior, permitió además utilizar las subrutinas ya desarrolladas.

El ejemplo de la tabla 16 muestra cuatro entradas con treinta y siete relaciones; después del proceso de reducción en la tabla 17 se muestra una sola entrada equivalente con las mismas treinta y siete relaciones; en el primer caso se requerían setenta y ocho cadenas para almacenar la información y en el segundo caso se requieren sólo cuarenta (una reducción del 49% sólo para esta entrada del diccionario).

Resumiendo, la reducción del diccionario se apoyó en la selección de las relaciones necesarias para el diccionario de escenarios; los tamaños se muestran en la tabla 18 donde se observa que EuroWordNet contiene en total 72,508 entradas (de relaciones individuales palabra-palabra) en el formato mostrado en la tabla 13; la primera selección redujo el número de entradas

a 27,959 con el formato de la tabla 16; y finalmente se obtuvieron 853 entradas en el formato de la tabla 17.

ENTRADA	RELACIONES
órgano	has_hyperonym boletín
órgano	has_hyperonym cosa
	has_hyponym retoño
	has_hyponym follaje
	has_hyponym cabillo
	has_hyponym offset
	has_hyponym raíz
	has_hyponym lámina
	has_hyponym estructura_reproductiva
	has_hyponym ascocarpo
	has_hyponym esporocarpo
órgano	has_hyperonym instrumento_de_viento
	has_mero_part pedal
	has_mero_part teclado
órgano	has_hyperonym parte_del_cuerpo
	has_hyponym órgano_eréctil
	has_hyponym órganos_reproductores
	has_hyponym centriolo
	has_hyponym condriosoma
	has_hyponym nucléolo
	has_hyponym núcleo
	has_hyponym cristalino
	has_hyponym órgano_del_habla
	has_hyponym lengua
	has_hyponym receptor
	has_hyponym víscera
	has_hyponym órgano_vital
	has_hyponym músculo
	has_hyponym ventosa
	has_hyponym patas
	has_hyponym oviscapto
	has_hyponym cilio
	has_mero_part lóbulo

Tabla 16 Ejemplo de entradas obtenidas de WordNet en Español

ENTRADA	RELACIONES
órgano	boletín cosa retoño follaje cabillo offset raíz lámina ascocarpo estructura_reproductiva esporocarpo instrumento_de_viento pedal teclado parte_del_cuerpo órgano_eréctil órganos_reproductores centriolo cristalino órgano_del_habla núcleo condriosoma nucléolo lengua receptor víscera órgano_vital músculo ventosa patas oviscapto cilio lóbulo

Tabla 17 Ejemplo de entradas del diccionario de escenarios

El proceso total de reducción, que se muestra en la tabla 18, permite observar una disminución drástica (alrededor del 98%) del almacenamiento necesario. Lo anterior se explica verificando el tamaño de cadenas de la tabla 16; las cadenas cosa, raíz, cilio, etc. (que permanecen en el diccionario) son cadenas de menor tamaño que *has_hyponym* (que se elimina); además de considerar la reducción del número de entradas en el diccionario al agrupar todas las relaciones en una sola entrada.

	TAMAÑO		
DESCRIPCIÓN	EuroWordNet	Relaciones	Final
Cadenas	217,488	91,917	3,210
Entradas	72,502	27,959	853
Tamaño (KB)	2,444	944	33

Tabla 18 Reducción del diccionario de escenarios