Capítulo 2

Recursos lingüísticos

Este capítulo está orientado a detallar los recursos lingüísticos que se han utilizado en el presente trabajo, la forma en la que éstos se presentan, y la manera en la que hemos hecho uso de ellos. Cada uno de esos recursos está asociado a un lenguaje natural concreto, es decir, a un idioma concreto. Los idiomas con los que más extensamente hemos trabajado han sido principalmente dos:

- El español, por ser de una utilidad práctica palpable. No sólo es el idioma más hablado en todo el territorio de nuestro país, sino también uno de los más hablados en todo el mundo.
- El inglés, por ser quizás el idioma que más ampliamente ha sido considerado por toda la comunidad científica internacional y, por tanto, por ser el idioma al que pertenecen los recursos lingüísticos de mayor calidad que se pueden encontrar.

Básicamente, consideraremos tres tipos de recursos:

- Textos etiquetados. Un texto o corpus etiquetado es aquél en el que todas y cada una de las palabras aparecen acompañadas al menos de su etiqueta correcta, es decir, aquélla que define correctamente el papel que dicha palabra está jugando en la frase concreta en la que aparece, y quizás también de su lema.

  El lema, o forma canónica del concepto al que hace referencia la palabra, es normalmente la forma del masculino singular, en el caso de sustantivos, adjetivos, etc., y el infinitivo en el caso de los verbos, es decir, lo que en definitiva constituye la entrada que podemos encontrar en cualquier diccionario semántico a la hora de buscar las acepciones o significados de un término dado.

  La principal aplicación de los *corpora* etiquetados es la de servir como referencia del comportamiento de un idioma, o al menos de un estilo de uso del mismo, durante el proceso de ajuste o puesta a punto de las herramientas destinadas a realizar cualquier tipo de procesamiento de nuevos textos en ese idioma, en particular su etiquetación.

- Diccionarios. En nuestro contexto, un diccionario, a veces diremos también *lexicón*, es simplemente una lista de palabras acompañada de sus posibles etiquetas o papeles candidatos que puede desempeñar dentro de la frase. Uno de nuestros objetivos concretos es comprobar de qué manera el uso de un diccionario externo puede ayudar a incrementar el rendimiento del proceso de etiquetación de textos en lenguaje natural.

- Bancos de árboles. Los bancos de árboles son textos donde no sólo cada palabra está acompañada de su etiqueta correcta, sino que además cada frase aparece completamente analizada sintácticamente. Los árboles de análisis con todos sus nodos se presentan en

línea de manera parentizada. Veremos cómo a partir de recursos de este tipo es posible extraer las reglas de la gramática a la que obedece un determinado idioma, o al menos un subconjunto relevante de él, incluso con una probabilidad de uso asignada a cada una de esas reglas.

Es obvio que un recurso de estas características presenta un nivel de anotación muy superior al de los textos simplemente etiquetados. Es por esta razón que intuimos que el uso de esta nueva información que aparece en él podría constituir una gran ayuda para el proceso de etiquetación, siendo quizás éste el principal objetivo que persigue el presente trabajo. Es evidente también que estos recursos requieren un esfuerzo de diseño mucho mayor, pero también es cierto que cada vez será más frecuente el poder encontrarlos disponibles, lo que nos lleva a pensar que el futuro de las técnicas que vamos a proponer no se verá comprometido.

A continuación presentamos en detalle todos y cada uno de los recursos lingüísticos que serán utilizados como fuentes de información para la formalización y descripción del comportamiento y características de los idiomas tratados.

2.1 El corpus ITU

Los recursos lingüísticos de calidad para el idioma español no son todavía demasiado abundantes. Uno de los textos ya etiquetados y libremente disponibles es el *International Telecommunications Union CCITT Handbook*, también conocido como *The Blue Book* en el ambiente de las telecomunicaciones, y como *Corpus ITU* en el ambiente lingüístico. Este corpus es la principal colección de textos sobre telecomunicaciones existente y, debido a su gran tamaño, constituye un excelente marco de pruebas para el estudio del comportamiento de los sistemas de etiquetación.

El texto original no etiquetado tiene alrededor de 5 millones de palabras. En lo que se refiere al texto ya anotado, existen dos versiones: el corpus entero etiquetado con la versión para español del etiquetador de XEROX [Cutting *et al.* 1992, Sánchez y Nieto 1995a, Sánchez y Nieto 1995b], y un subcorpus de aproximadamente medio millón de palabras corregido a mano por la Universidad de Lancaster. Esta segunda versión es la que se ha utilizado como uno de los *corpora* de referencia del presente trabajo, aunque ambos se pueden descargar desde las URL,s[1] correspondientes a las páginas que soportan toda la información relativa al proyecto CRATER: *Corpus Resources And Terminology ExtRaction* [CRATER 1993].

A continuación describimos detalladamente las principales características de nuestro corpus de referencia:

- El corpus tiene 486.073 palabras. El tamaño del fichero es de 7.564.781 caracteres. La sintaxis de cada línea es de la forma

  ```
  palabra/etiqueta/lema  palabra/etiqueta/lema ... palabra/etiqueta/lema
  ```

  Cada una de las líneas del fichero es una frase. El corpus tiene 14.919 frases, es decir, un número medio de 33 palabras por frase.

- El corpus contiene 45.679 formas verbales, donde 581 formas están en voz pasiva, 870 son formas verbales compuestas, y 3.415 son formas verbales con un pronombre enclítico.

- Las características del lexicón o diccionario constituido por todas las formas que aparecen en el corpus son las siguientes:

[1] *Uniform Resource Locator* (dirección de un recurso de Internet).

15.745 formas con 1 etiqueta, 1.097 formas con 2 etiquetas,
223 formas con 3 etiquetas, 61 formas con 4 etiquetas,
8 formas con 5 etiquetas, 3 formas con 6 etiquetas y
1 forma con 7 etiquetas.

Esto es, 17.138 formas diferentes, con 18.917 etiquetas posibles, y correspondientes a 12.462 lemas diferentes. Si calculamos el porcentaje de formas ambiguas y el número medio de etiquetas por forma, obtenemos:

$$\%\text{ formas ambiguas} = \frac{\#\text{ formas ambiguas}}{\#\text{ formas}} \times 100 = \frac{17.138 - 15.745}{17.138} \times 100 = 8,13\ \%$$

$$\#\text{ medio de etiquetas por forma} = \frac{\#\text{ etiquetas}}{\#\text{ formas}} = \frac{18.917}{17.138} = 1,10 \text{ etiquetas por forma.}$$

- Mucho más interesante es calcular las mismas características directamente con todas las palabras del corpus, y obtenemos las siguientes cifras:

  263.592 palabras con 1 etiqueta, 107.746 palabras con 2 etiquetas,
  79.751 palabras con 3 etiquetas, 7.013 palabras con 4 etiquetas,
  18.949 palabras con 5 etiquetas, 8.528 palabras con 6 etiquetas y
  494 palabras con 7 etiquetas.

  Esto es, 486.073 palabras, con 895.760 etiquetas posibles. Si calculamos de nuevo el porcentaje de palabras ambiguas y el número medio de etiquetas por palabra, obtenemos:

$$\%\text{ palabras ambiguas} = \frac{\#\text{ palabras ambiguas}}{\#\text{ palabras}} \times 100 = \frac{486.073 - 263.592}{486.073} \times 100 = 45,77\ \%$$

$$\#\text{ medio de etiquetas por palabra} = \frac{\#\text{ etiquetas}}{\#\text{ palabras}} = \frac{895.760}{486.073} = 1,84 \text{ etiquetas por palabra.}$$

Ejemplo 2.1 A continuación mostramos una línea tomada directamente del corpus ITU, para ilustrar el aspecto general que presentan los datos:

```
En/PREP/en esta/DMPXFS/este colaboraci&oacute;n/NCFS/colaboraci&oacute;n
debe/VMPI3S/deber reconocerse/VCLI/reconocer el/ARTDMS/el
car&aacute;cter/NCMS/car&aacute;cter consultivo/ADJGMS/consultivo de/PREP/de
las/ARTDFP/el organizaciones/NCFP/organizaci&oacute;n que/CQUE/que
participan/VLPI3P/participar en/PREP/en los/ARTDMP/el trabajos/NCMP/trabajo
del/PDEL/del CCITT/NPTOS/CCITT ,/CM/, en_particular/ADVN/en_particular
a/PREP/a la/ARTDFS/el ISO/ACRNM/ISO ,/CM/, desde/PREP/desde el/ARTDMS/el
punto/NCMS/punto de/PREP/de vista/NCFS/vista de/PREP/de su/PPOSPS/suyo
labor/NCFS/labor con_respecto_a/PREP/con_respecto_a los/ARTDMP/el
sistemas/NCMP/sistema de/PREP/de datos/NCMP/dato y/CC/y a/PREP/a
las/ARTDFP/el comunicaciones/NCFP/comunicaci&oacute;n ./FS/.
```

□

Las etiquetas que aparecen a lo largo de este corpus pertenecen al juego de etiquetas desarrollado en el marco del proyecto CRATER, el cual obedece al estándar EAGLES[2]. La descripción de cada una de ellas se puede ver en la sección A.1.

2.2 El sistema GALENA

Tal y como ya se ha explicado en la introducción de la presente memoria, uno de los objetivos que perseguimos en este trabajo es la evaluación y comparación de distintos sistemas de etiquetación, y el estudio de sus comportamientos concretos cuando manejan textos en español. Entre esos sistemas a comparar figura nuestra propia herramienta: el sistema GALENA [Vilares *et al.* 1995]. Durante las fases de diseño e implementación del analizador léxico de dicha herramienta, se ha desarrollado un lexicón de tamaño medio para el español [Graña *et al.* 1994], el cual nos ha parecido interesante integrar en los distintos sistemas de etiquetación a evaluar, siempre que esto ha sido posible, para observar en qué medida un diccionario externo puede ayudar a mejorar el proceso de etiquetación en cada uno de los sistemas.

En este punto, para poder llevar a cabo correctamente esa integración, surge el problema de que el corpus ITU y el diccionario del sistema GALENA no utilizan el mismo juego de etiquetas. Para solventar esta dificultad es necesario establecer primero una correspondencia entre ambos juegos de etiquetas, el del proyecto CRATER y el del sistema GALENA, y después existen dos posibilidades:

1. Transformar las etiquetas de todas las palabras del corpus ITU en etiquetas del sistema GALENA.

2. Transformar las etiquetas de todas las palabras del diccionario del sistema GALENA en etiquetas CRATER.

Hemos optado por la primera de las posibilidades, por ser la más sencilla y por ser la que en principio no altera los recursos que nosotros mismos hemos generado y que en el fondo son los que más nos interesa contrastar.

Por tanto, el juego de etiquetas del sistema GALENA es el que ha sido utilizado en todos los experimentos relativos al idioma español que se han realizado en el presente trabajo. Así pues, las siguientes secciones describen detalladamente las características de los dos principales recursos lingüísticos aportados por el sistema GALENA: su juego de etiquetas y su diccionario.

2.2.1 El juego de etiquetas del sistema GALENA

La tabla 2.1 es una representación compacta del juego de etiquetas del sistema GALENA. Los campos están en la fila en negrita. Los valores de `Categoría` y `Tipo` están hacia abajo. Los valores para el resto de los campos están hacia arriba. Los cruces en el resto de la tabla muestran los valores permitidos para cada caso. Por ejemplo, `Grado` no tiene sentido para un `Verbo`, y sólo `singular` y `plural` están permitidos para el `Número persona` de un `Posesivo`. El género tiene sentido en los verbos cuando se trata de una forma verbal en participio y, en todo caso, dicha marca de género es el último carácter de las etiquetas de los verbos.

[2] *Expert Advisory Group on Language Engineering Standards* (EAGLES) es una iniciativa de la Comisión Europea iniciada en 1993 dentro del marco del programa de Investigación e Ingeniería Lingüística (LRE), y su objetivo es acelerar la provisión de estándares para la construcción de recursos lingüísticos a gran escala (*corpora* de textos y de habla, léxicos informatizados, etc.), así como para la manipulación de este conocimiento (formalismos lingüísticos, lenguajes de marcado y herramientas informáticas), y establecer mecanismos de evaluación de los recursos, herramientas y productos [Calzolari y McNaught 1996].

Tabla 2.1: Juego de etiquetas del sistema GALENA

Categoría	Tipo	Subtipo	Género	Número	Grado	Persona	Número persona	Caso	Tiempo verbal	Modo
Sustantivo (S)	Común (c)		m f y	s p y						
	Propio (p)		m f y	s p y						
Adjetivo (A)			m f y	s p y	c 0					
Demostrativo (E)		n y	m f n y	s p y						
Relativo (T)		n d y	m f n y	s p y						
Indefinido (I)		n d y	m f n y	s p y						
Interrogativo (G)		n y	m f n y	s p y						
Posesivo (M)		n d y	m f n y	s p y		1 2 3	s p			
Numeral (N)	Cardinal (c)	n d y	m f y	s p						
	Ordinal (o)	n d y	m f y	s p						
	Partitivo (p)	n	m f y	s p						
	Múltiplo (m)	n	m f y	s p						
Artículo (D)			m f n	s p						
Pronombre Personal (R)	Tónico (t)		m f y (último)			1 2 3	s p y	n a d y p q		
	Proclítico átono (p)		m f y (último)			1 2 3	s p y	n a d y p q		
	Enclítico átono (e)		m f y (último)			1 2 3	s p y	n a d y p q		
Verbo (V)			m f 0 (último)			1 2 3 y 0	s p 0		p i s e f c 0 P I S E F C	i s m f F g G p
Preposición (P)										
Conjunción (C)	Coordinada (c)									
	Subordinada (s)									
Adverbio (W)	Nuclear (n)									
	Modificador (m)									
	Nuclear y modificador (y)									
	Relativo (r)									
	Exclamativo (i)									
	Interrogativo (v)									
Interjección (Y)										
Marca Puntuación (Q)	. , ; : ¿ ? ¡ ! ( ) - " ...									
Periférica (Z)	Palabra extranjera (e)		m f y 0	s p y 0						
	Fórmula (f)		m f y 0	s p y 0						
	Abreviatura (a)		m f y 0	s p y 0						
	Símbolo (s)		m f y 0	s p y 0						
	Sigla (g)		m f y 0	s p y 0						
	Otros (o)		m f y 0	s p y 0						

Categoría: Tipo; Subtipo: Det. y no det. (y), Determinante (d), No determinante (n); Género: Masc. y fem. (y), Neutro (n), Femenino (f), Masculino (m); Número: No aplicable (0), Singular y plural (y), Plural (p), Singular (s); Grado: No aplicable (0), Comparativo (c); Persona: No aplicable (0), Tercera (3), Segunda (2), Primera (1); Número persona: No aplicable (0), Primera y tercera (y), Singular y plural (y), Plural (p), Singular (s); Caso: No aplicable (0), Acusativo y dativo (y), Dativo (d), Acusativo (a), Nominativo (n); Tiempo verbal: Nominativo y caso con preposición (q), Caso con preposición (p), Copretérito_se (s), Copretérito (i), Pretérito (e), Presente (p); Modo: No aplicable (0), Antepostpretérito (C), Antefuturo (F), Antecopretérito_se (S), Antecopretérito (I), Antepretérito (E), Antepresente (P), Postpretérito (c), Futuro (f), Participio (p), Gerundio compuesto (G), Gerundio (g), Infinitivo compuesto (F), Infinitivo (f), Imperativo (m), Subjuntivo (s), Indicativo (i)

Ejemplo 2.2 Las tres etiquetas de la palabra `sobre` serían: `Scms` (sustantivo, común, masculino, singular), `P` (preposición), y `Vysps0` (verbo, primera y tercera personas, singular, presente, subjuntivo, género no aplicable). □

Esta tabla genera un conjunto de 1.048 etiquetas, aunque no todas ellas representan combinaciones correctas desde un punto de vista lingüístico. Por otro lado, para el presente estudio, hemos decidido extender el conjunto de etiquetas del sistema GALENA con el fin de marcar explícitamente las formas verbales compuestas, las formas verbales en voz pasiva y las formas verbales con pronombre enclíticos. Por tanto, el conjunto final de etiquetas que hemos utilizado consta de las 373 etiquetas que aparecen en la sección A.2.

A continuación describimos el procedimiento que se ha seguido para extender la marcación original de las formas verbales en el sistema GALENA. Básicamente, si `<etiqueta>` es la etiqueta que aparecía originalmente, la nueva marcación consiste en añadir un sufijo a esa `<etiqueta>`. En el caso de las formas verbales compuestas, tales como `he comido`, aparecen dos componentes:

1. El verbo auxiliar (`he`), que será marcado como `<etiqueta>TC1`.
2. El participio (`comido`), que será marcado como `<etiqueta>TC2`.

Para las formas verbales en voz pasiva, tales como `fue comido`, aparecen dos componentes:

1. El verbo auxiliar (`fue`), que será marcado como `<etiqueta>TCP1`.
2. El participio (`comido`), que será marcado como `<etiqueta>TCP2`.

O tres, en casos como `ha sido comido`:

1. El primer verbo auxiliar (`ha`), que será marcado como `<etiqueta>TCP1`.
2. El segundo verbo auxiliar (`sido`), que será marcado como `<etiqueta>TCP2`.
3. El participio (`comido`), que será marcado como `<etiqueta>TCP3`.

Hasta aquí, este procedimiento presenta las siguientes implicaciones:

- Cada una de las formas del verbo `haber` puede tener tres etiquetaciones diferentes:
  - `<etiqueta>`, cuando va solo.
  - `<etiqueta>TC1`, cuando es la primera parte de un tiempo compuesto.
  - `<etiqueta>TCP1`, cuando es la primera parte de una voz pasiva.
- Cada una de las formas del verbo `ser` puede tener dos etiquetaciones diferentes:
  - `<etiqueta>`, cuando va solo.
  - `<etiqueta>TCP1`, cuando es la primera parte de una voz pasiva.
- Cada participio verbal puede tener hasta cinco etiquetaciones diferentes:
  - `V0s0pm`, cuando es participio.
  - `V0s0pmTC2`, cuando es la segunda parte de un tiempo compuesto.
  - `V0s0pmTCP2`, cuando es la segunda parte de una voz pasiva.
  - `V0s0pmTCP3`, cuando es la tercera parte de una voz pasiva.
  - `Ams0`, cuando es adjetivo.

Ejemplo 2.3 Las cinco etiquetaciones anteriores son las del participio masculino singular. Más exactamente, por ejemplo para el participio del verbo `ver`, tenemos todas estas formas y etiquetaciones posibles:

```
visto V0s0pm V0s0pmTC2 V0s0pmTCP2 V0s0pmTCP3 Ams0
vista V0s0pf V0s0pfTCP2 V0s0pfTCP3 Afs0
vistos V0p0pm V0p0pmTCP2 V0p0pmTCP3 Amp0
vistas V0p0pf V0p0pfTCP2 V0p0pfTCP3 Afp0
```
□

Respecto a las formas verbales con pronombres enclíticos, tales como `tenerlo` o `verse`, en español pueden tener hasta tres de esos pronombres, como en `tráetemelo`. Sin embargo, el corpus ITU presenta formas con sólo un pronombre enclítico, que serán marcadas como `<etiqueta>PE1`.

Por último, sólo nos falta indicar cómo es exactamente la correspondencia entre las etiquetas CRATER y las etiquetas GALENA. El juego de etiquetas del proyecto CRATER resulta ser un poco más preciso que el del sistema GALENA, debido a que tiene un cardinal superior: 475 etiquetas frente a 373. A pesar de esto, dicha correspondencia se puede establecer sin problema alguno, tal y como se especifica en la sección A.3. La aplicación de esta correspondencia al corpus ITU nos ha proporcionado un corpus de referencia para nuestro experimento completamente etiquetado con el juego de etiquetas del sistema GALENA.

Ejemplo 2.4 A continuación reproducimos de nuevo la misma porción de texto que habíamos utilizado en el ejemplo 2.1 para ilustrar el aspecto general del corpus, pero esta vez ya con las etiquetas definitivas:

```
En/P/en esta/Eyfs/este colaboración/Scfs/colaboración debe/V3spi0/deber
reconocerse/V000f0PE1/reconocer el/Dms/el carácter/Scms/carácter
consultivo/Ams0/consultivo de/P/de las/Dfp/el
organizaciones/Scfp/organización que/Cs/que participan/V3ppi0/participar
en/P/en los/Dmp/el trabajos/Scmp/trabajo de/P/de el/Dms/el CCITT/Spys/CCITT
,/Q,/, en_particular/Wy/en_particular a/P/a la/Dfs/el ISO/Zgfs/ISO ,/Q,/,
desde/P/desde el/Dms/el punto/Scms/punto de/P/de vista/Scfs/vista de/P/de
su/Mdys3s/suyo labor/Scfs/labor con_respecto_a/P/con_respecto_a los/Dmp/el
sistemas/Scmp/sistema de/P/de datos/Scmp/dato y/Cc/y a/P/a las/Dfp/el
comunicaciones/Scfp/comunicación ./Q./.
```
□

Como se puede ver, no sólo se han efectuado cambios al nivel de las etiquetas, sino también al nivel de las palabras y los lemas[3], para adaptarlas y hacerlas totalmente compatibles con las entradas del diccionario del sistema GALENA, el cual se describe a continuación.

2.2.2 El diccionario del sistema GALENA

Como hemos visto anteriormente, el segundo de los recursos lingüísticos importantes que aporta el sistema GALENA es su diccionario, un lexicón de tamaño medio desarrollado durante las fases de diseño e implementación del analizador léxico de dicho sistema.

Sin embargo, antes de describir directamente este recurso, vamos a ver algunas cifras correpondientes a las características generales del sistema GALENA, en lo que al análisis léxico se refiere:

[3] Cabe destacar, por ejemplo, la presencia de acentos reales.

- La base de datos léxica contiene el siguiente número de raíces:

2.944 adjetivos,	125 adverbios,	2 artículos,	102 conjunciones,
14 demostrativos,	36 indefinidos,	40 interjecciones,	4 interrogativos,
62 numerales,	67 periféricas,	8 posesivos,	43 preposiciones,
29 pronombres,	5 relativos,	8.027 sustantivos,	5.600 verbos y
30 especiales.			

  Esto es, 17.138 raíces correspondientes a 13.667 lemas diferentes.

- Existe un proceso de compilación que transforma la base de datos léxica en una arquitectura de reconocimiento mucho más eficiente basada en un autómata finito acíclico determinista numerado[4], el cual consta de 11.985 estados y 31.258 transiciones.

- El tamaño del fichero compilado correspondiente a ese autómata finito es de 3.466.121 *bytes*.

- El tiempo de compilación es de aproximadamente 7 segundos, y la velocidad de reconocimiento de 40.000 palabras por segundo en una máquina con un procesador Pentium II a 300 MHz bajo sistema operativo Linux.

Para aquellas herramientas de etiquetación en las que sea posible la integración de un diccionario externo, y a partir de la base de datos léxica anteriormente descrita, hemos generado todas las formas reconocidas por el analizador léxico del sistema GALENA, y las características de este diccionario han resultado ser las siguientes:

265.418 formas con 1 etiqueta,	9.995 formas con 2 etiquetas,
588 formas con 3 etiquetas,	11.343 formas con 4 etiquetas,
4.097 formas con 5 etiquetas y	163 formas con 6 etiquetas.

Es decir, 291.604 formas diferentes, con 354.007 etiquetas posibles, y correspondientes, como ya hemos dicho, a 13.667 lemas diferentes.

Sin embargo, la base de datos léxica no almacena ninguna información sobre frecuencias o probabilidades, ya que esta información es relevante para las palabras, no para las raíces, y además depende de cada corpus concreto. Por tanto, si utilizamos únicamente la base de datos léxica, el diccionario generado contendrá las etiquetas candidatas de cada forma en un orden totalmente arbitrario y no resultará fiable para determinados sistemas de etiquetación, como puede ser el caso del etiquetador de Brill, el cual considera la primera etiqueta de cada palabra como la más probable [Brill 1992]. Para evitar este problema, hemos realizado un estudio de las clases de ambigüedad presentes en el corpus ITU, y hemos generado el diccionario correspondiente al sistema GALENA con las etiquetas de cada forma ordenadas según la combinación de etiquetas más frecuente en el corpus ITU para su clase de ambigüedad.

Para dar una idea aproximada de lo que este diccionario externo puede aportar, diremos que la intersección del diccionario del sistema GALENA y el diccionario constituido por todas las formas que aparecen en el corpus ITU contiene 6.594 formas, donde:

- 3.670 formas (que involucran a 166.573 palabras del corpus ITU, es decir, a un 34,27% de las palabras) están en la misma clase de ambigüedad en ambos diccionarios.

- 2.924 formas (que involucran a 216.106 palabras del corpus ITU, es decir, a un 44,46% de las palabras) están en diferentes clases de ambigüedad en ambos diccionarios.

[4] El método de construcción de este tipo autómatas se describe detalladamente en el capítulo 3.

Para no apartar demasiado las condiciones de nuestros experimentos de las que se producen realmente en la práctica, las etiquetas de todas estas formas en común, incluso de las que están en la misma clase de ambigüedad, han sido generadas con el orden obtenido al aplicar el mismo método general que hemos explicado anteriormente. Es cierto que podríamos haberlas generado con las etiquetas ordenadas exactamente como están en el diccionario constituido por todas las formas que aparecen en el corpus ITU, pero, como ya hemos esbozado anteriormente, disponer de un corpus sobre el cual se pueda realizar un estudio previo sobre el léxico o sobre las clases de ambigüedad no es lo usual.

2.3 El corpus SUSANNE

El corpus SUSANNE [Sampson 1994a] es el banco de árboles que se ha utilizado en nuestro estudio como corpus de referencia para el idioma inglés. El corpus SUSANNE se ha creado con el apoyo del *Economic and Social Research Council (UK)*, como parte del proceso de desarrollo de una taxonomía exhaustiva orientada al NLP y de un esquema para la grámatica del inglés, tanto a nivel lógico como de superficie: el esquema SUSANNE.

El esquema SUSANNE [Sampson 1994b] intenta proporcionar un método de representación de todos los aspectos de la gramática del inglés que están suficientemente bien definidos como para ser subceptibles de anotarse formalmente. El modelo ofrece un esquema de categorías y un conjunto de formas de aplicarlas que lo hacen muy práctico para los investigadores en NLP a la hora de registrar sistemáticamente y sin ambigüedades todo lo que ocurre en el uso real del lenguaje, y permite que investigadores de diferentes lugares puedan intercambiar datos gramaticales sin que surjan confusiones relacionadas con los usos y terminologías locales.

El corpus SUSANNE se ha producido casi en su totalidad de forma manual, no a través de un sistema de análisis sintáctico automático. El corpus en sí mismo consiste en un subconjunto de aproximadamente 150.000 palabras (correspondientes a unos 10.500 lemas diferentes) del corpus BROWN para el inglés americano [Francis y Kučera 1982], anotado de acuerdo con el esquema SUSANNE. Por tanto, el esquema SUSANNE para la anotación de los análisis sintácticos se ha desarrollado en base al inglés británico y al americano. No cubre otros lenguajes, aunque se espera que sus principios generales sirvan de ayuda al desarrollo de otras taxonomías comparables para ellos. El esquema está principalmente orientado al lenguaje escrito.

2.3.1 Estructura del corpus SUSANNE

El corpus SUSANNE consta de 64 ficheros, cada uno de los cuales contiene una versión anotada de un texto de unas 2.000 palabras del corpus BROWN. Los ficheros tienen un tamaño medio de unos 83 *kilobytes* y, por tanto, el corpus completo ocupa unos 5.3 *megabytes*. Los nombres de los ficheros son los mismos que los de los respectivos textos BROWN, por ejemplo `A01`, `N12`, etc. Se han analizado 16 textos de cada una de las cuatro categorías o géneros literarios BROWN siguientes:

- `A`: reportajes de prensa.
- `G`: bellas letras, biografías, memorias.
- `J`: textos eruditos, principalmente científicos y técnicos.
- `N`: aventuras y *Western* de ficción.

Cada fichero tiene una línea por cada palabra del texto original. Sin embargo, las *palabras* en el corpus SUSANNE son a menudo más pequeñas que las palabras en el sentido ortográfico

ordinario. Por ejemplo, las marcas de puntuación y los sufijos de apóstrofo `s` se tratan como palabras separadas y se les asigna líneas distintas. Por otra parte, allí donde los caracteres no se pueden representar adecuadamente mediante el uso directo del juego de caracteres estándar, éstos se representan mediante entidades cuyos nombres figuran entre ángulos. Cuando ha sido posible, esos nombres se han tomado del SGML[5]. Por ejemplo, `<eacute>` es el símbolo utilizado para una `e` minúscula con acento. Los símbolos entre ángulos se utilizan también para representar situaciones tales como los cambios tipográficos, los cuales también se representan dentro de la secuencia de palabras como elementos separados. Por ejemplo, `<bital>` significa *begin italics*, es decir, comienzo de una porción de texto en letra itálica.

Cada línea de un fichero SUSANNE consta de seis campos separados por tabuladores. Cada campo de cada línea contiene al menos un caracter. Los seis campos de cada línea son los siguientes:

1. Campo de referencia. El campo de referencia contiene 9 caracteres los cuales dan a cada línea un número de referencia que es único a lo largo de todo el corpus SUSANNE, por ejemplo, `N06:1530t`. Los primeros tres caracteres (`N06`, en el ejemplo anterior) son el nombre del fichero. El cuarto caracter es siempre el caracter de dos puntos. Los caracteres quinto al octavo (en el ejemplo `1530`) son el número de la línea en la versión *Bergen I* del corpus BROWN en la cual aparece la palabra. Los números de línea en el corpus BROWN se incrementan normalmente en diez unidades, apareciendo ocasionalmente otros números no múltiplos de diez intercalados. Y el noveno caracter es una letra minúscula que diferencia las palabras sucesivas que aparecen en la misma línea BROWN. Las líneas SUSANNE se enumeran de manera continua desde la `a`, omitiendo la `l` y la `o`.

2. Campo de estado. El campo de estado consta de un caracter. Las letras `A` y `S` indican que la palabra es una abreviatura o un símbolo, respectivamente, tal y como están definidos los códigos del corpus BROWN.

   El corpus SUSANNE está orientado a reflejar la incidencia de los errores que aparecen en el inglés escrito de la vida real y, por tanto, a reproducir esos errores tal y como aparecen en los textos originales en los que se basa el corpus BROWN, pero corrigiendo los errores que fueron introducidos durante el proceso de la construcción del corpus. Cuando un error, o un error aparente, refleja la forma encontrada en la publicación original, se conserva en el corpus SUSANNE marcada con una `E` en el campo de estado. En otro caso, el corpus SUSANNE recupera el texto de la publicación original y el campo de estado ignora el error. Cuando los errores son originales, la etiquetación de las palabras y el análisis gramatical se aplican al texto erróneo de la mejor manera posible estableciendo analogías con las formas correctas.

   En la gran mayoría de las líneas, no se aplica ninguna de estas tres categorías y el campo de estado contiene simplemente un guión.

3. Campo para la etiqueta de la palabra. El conjunto de etiquetas de las palabras en el corpus SUSANNE está basado en el juego de etiquetas LANCASTER [Garside *et al.* 1987]. No obstante, en este juego de etiquetas se han efectuado algunas distinciones gramaticales adicionales indicadas mediante letras en minúsculas que se añaden como sufijos a las etiquetas LANCASTER[6]. Aparte de estas extensiones en minúsculas, las etiquetas son normalmente idénticas a las etiquetas LANCASTER. A las marcas de puntuación se les

[5] *Standard Generalized Markup Language* es un metalenguaje para la marcación o codificación electrónica de textos, adoptado como estándar internacional en 1986 (ISO 8879) [Burnard 1995].

[6] Por ejemplo, `revealing` se etiqueta como `VVG` (participio presente de un verbo) en el esquema LANCASTER, pero como `VVGt` (participio presente de un verbo transitivo) en el esquema SUSANNE.

asignan etiquetas alfabéticas que comienzan por `Y`[7], y el signo del dólar que aparece en algunas etiquetas LANCASTER para el genitivo se reemplaza por `G`[8], de manera que las etiquetas LANCASTER modificadas siempre contienen caracteres alfanuméricos y comienzan con letras mayúsculas.

La etiqueta `YG` aparece en este campo para representar un elemento *fantasma* o una *traza*, es decir, la posición lógica de un constituyente que en la estructura gramatical de superficie ha sido eliminado o desplazado a otra posición.

El juego de etiquetas SUSANNE contiene 425 etiquetas distintas, y se muestra en la sección A.4.

4. Campo para la palabra. Este campo contiene un segmento de texto que normalmente coincide con una palabra en el sentido ortográfico, pero que a veces, tal y como hemos indicado anteriormente, incluye sólo parte de esa palabra ortográfica. Hemos visto también que, en general, este campo representa los mismos fenómenos tipográficos que aparecen en el corpus BROWN, aunque en determinados casos el corpus SUSANNE ha considerado los documentos originales con el fin de reconstruir detalles tipográficos omitidos por el corpus BROWN.

   Algunos caracteres tienen significados especiales en este campo:

   - `+` aparece sólo como primer caracter del campo de la palabra para indicar que en el texto original el contenido del campo no estaba separado del segmento de texto inmediatamente precedente mediante espacio en blanco[9].
   - `-` indica que la línea no corresponde a nigún texto material, sino que representa la *traza* de un elemento que ha sido movido.
   - `<...>` encierran nombres de entidad para características tipográficas especiales, tal y como se discutió anteriormente.

5. Campo para el lema. El campo para el lema muestra la entrada de diccionario de la cual la palabra del texto es una forma, es decir, muestra las formas canónicas de las palabras que aparecen flexionadas en el texto, y también elimina las variaciones tipográficas, tales como la primera letra mayúscula, que no son inherentes a la palabra, sino al contexto en el que se usa. En el caso de las palabras para las cuales el concepto de entrada de diccionario es inapropiado, por ejemplo los numerales y las marcas de puntuación, el campo del lema contiene un guión.

6. Campo para el análisis sintáctico. El contenido de este sexto campo representa la información central del corpus SUSANNE. En él se codifica la estructura gramatical de los textos como una secuencia de árboles etiquetados, los cuales tienen un nodo hoja para cada una de las líneas del corpus.

   Cada texto se trata como una secuencia de párrafos separados por cabeceras. Un párrafo coincide normalmente con un párrafo ortográfico ordinario. Una cabecera puede constar de un texto material real, o puede ser una mera división tipográfica de párrafo, simbolizada mediante `<minbrk>` en el campo de la palabra. Conceptualmente, la estructura de cada párrafo o cabecera es un árbol con un nodo raíz etiquetado como `O` (`Oh` para una cabecera), y con un nodo hoja etiquetado con una etiqueta de palabra para cada palabra SUSANNE

[7] Por ejemplo, `YC` para la coma.

[8] Por ejemplo, `GG` para el sufijo de apóstrofo `s`.

[9] Por ejemplo, en el caso de una marca de puntuación, o en el caso de una secuencia con guiones que ha sido rota en varias líneas SUSANNE.

o para cada traza, es decir, para cada línea del corpus. Comúnmente existirán muchos nodos intermedios también etiquetados.

Estos árboles se representan mediante cadenas de caracteres parentizadas o con corchetes en la forma usual, con las etiquetas de los nodos no terminales escritas dentro de los dos corchetes, es decir, a la derecha del corchete de apertura y a la izquierda del corchete de cierre. Esta cadena se adapta como sigue para incluir en ella sucesivos campos de análisis. Cada vez que un corchete de apertura sigue a un corchete de cierre, la cadena se segmenta produciéndose un segmento por cada nodo hoja. Y dentro de cada uno de esos segmentos, la secuencia `[ etiqueta ]` que representa al nodo hoja se sustituye por un punto. Por tanto, cada campo de análisis contiene exactamente un punto, que corresponde al nodo terminal etiquetado con el contenido del campo 3, el campo para la etiqueta de la palabra, a veces precedido por etiquetas de nodos y corchetes de apertura y a veces seguido por etiquetas de nodos y corchetes de cierre, los cuales correponden a los subárboles que empiezan o finalizan con la palabra de la línea en cuestión.

En total los árboles de análisis del corpus SUSANNE contienen 267.046 nodos, de los cuales 4.383 son raíces y 156.584 son hojas. La notación para las etiquetas de los nodos no terminales se muestra en el apéndice B.

Ejemplo 2.5 Para ilustrar el aspecto general que presentan los datos, a continuación mostramos un conjunto de líneas tomadas directamente del corpus SUSANNE:

```
...
A01:0010b  -  AT     The            the            [O[S[Nns:s.
A01:0010c  -  NP1s   Fulton         Fulton         [Nns.
A01:0010d  -  NNL1cb County         county         .Nns]
A01:0010e  -  JJ     Grand          grand          .
A01:0010f  -  NN1c   Jury           jury           .Nns:s]
A01:0010g  -  VVDv   said           say            [Vd.Vd]
A01:0010h  -  NPD1   Friday         Friday         [Nns:t.Nns:t]
A01:0010i  -  AT1    an             an             [Fn:o[Ns:s.
A01:0010j  -  NN1n   investigation  investigation  .
A01:0020a  -  IO     of             of             [Po.
A01:0020b  -  NP1t   Atlanta        Atlanta        [Ns[G[Nns.Nns]
A01:0020c  -  GG     +<apos>s       -              .G]
A01:0020d  -  JJ     recent         recent         .
A01:0020e  -  JJ     primary        primary        .
A01:0020f  -  NN1n   election       election       .Ns]Po]Ns:s]
A01:0020g  -  VVDv   produced       produce        [Vd.Vd]
A01:0020h  -  YIL    <ldquo>        -              .
A01:0020i  -  ATn    +no            no             [Ns:o.
A01:0020j  -  NN1u   evidence       evidence       .
A01:0020k  -  YIR    +<rdquo>       -              .
A01:0020m  -  CST    that           that           [Fn.
A01:0030a  -  DDy    any            any            [Np:s.
A01:0030b  -  NN2    irregularities irregularity   .Np:s]
A01:0030c  -  VVDv   took           take           [Vd.Vd]
A01:0030d  -  NNL1c  place          place          [Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
A01:0030e  -  YF     +.             -              .O]
...
```

Este conjunto de líneas constituyen una frase completa. El sexto campo de todas estas líneas es una representación parentizada del árbol de análisis sintáctico que se muestra gráficamente en la figura 2.1. □

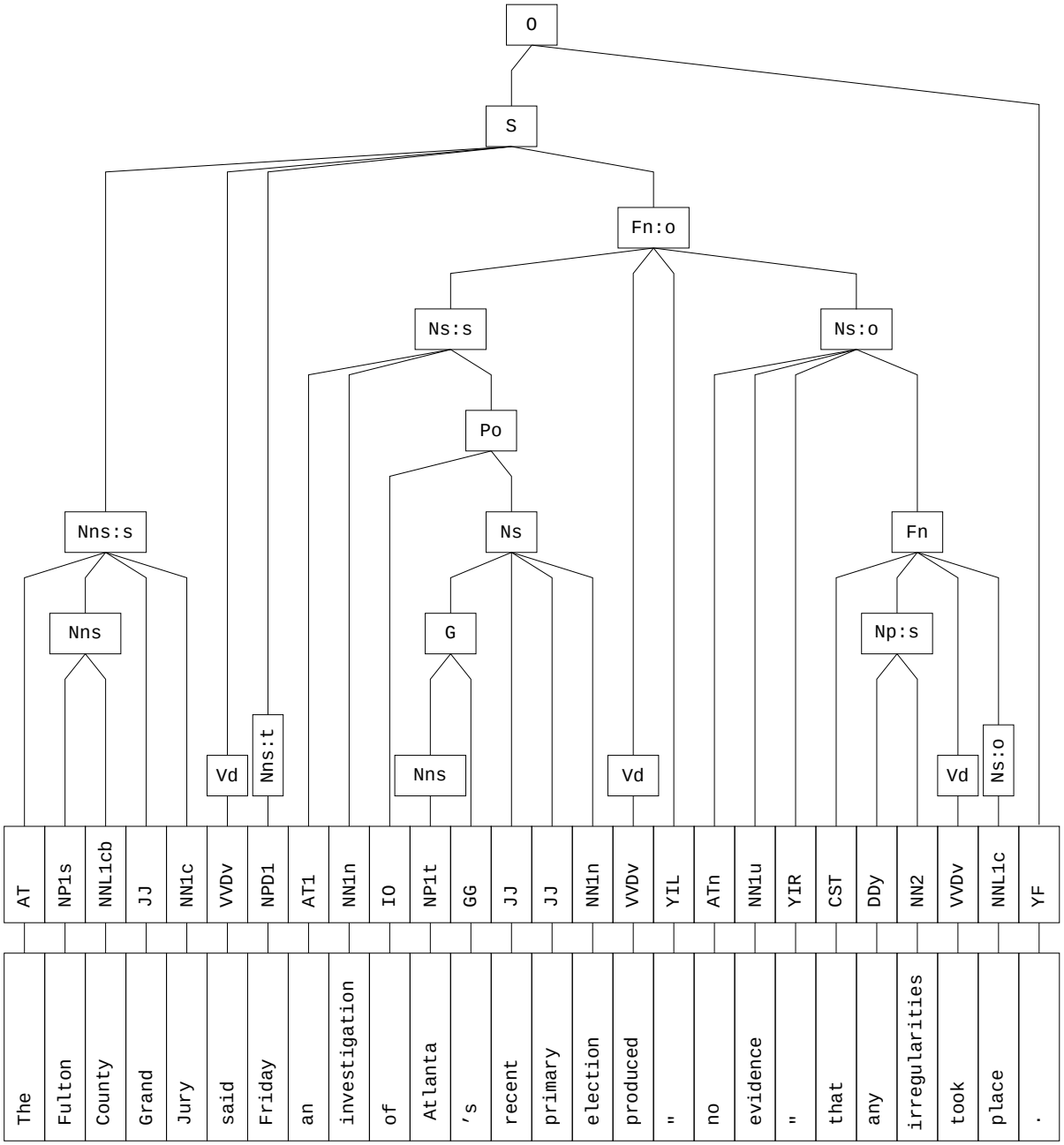

Figura 2.1: Ejemplo de árbol sintáctico para una frase del corpus SUSANNE

La siguiente sección está destinada a describir las características de los recursos lingüísticos que hemos extraído a partir del corpus SUSANNE: una gramática probabilística y dos *corpora* etiquetados.

2.3.2 Extracción de recursos del corpus SUSANNE

La primera transformación real que hemos realizado con el corpus SUSANNE ha sido integrar en una sola línea toda la información léxica y sintáctica disponible para cada frase, manteniedo en los análisis la notación ya introducida, esto es, `[` y `]` cada vez que comienza y finaliza un subárbol, respectivamente. Sin embargo, como podemos ver en el apéndice B, el corpus SUSANNE contiene símbolos no terminales para estructuras de nivel más alto que la frase, como por ejemplo párrafos y títulos. Por tanto, hemos tomado cada análisis y hemos eliminado todo lo innecesario, por ejemplo, las marcas de párrafos y las de títulos, para mantener sólo los subanálisis que corresponden a frases reales, es decir, aquéllos cuyo símbolo no terminal raíz comienza con `S`, `F`, `T`, `W`, `A`, `Z` o `L`.

2.3.2.1 Una gramática probabilística

Nuestro principal objetivo ahora es el de extraer una gramática para realizar nuestros experimentos sobre el corpus SUSANNE.

Como ya hemos visto también, el corpus SUSANNE contiene dos tipos de frases: con trazas y sin trazas. Las frases con trazas pueden ser útiles, por ejemplo, para estudiar otro tipo de fenómenos lingüísticos en los cuales se hace referencia de una manera no explícita a otros componentes de la frase, o incluso de otras frases, tales como la anáfora o la catáfora. Sin embargo, en nuestro caso, si hubiésemos utilizado este tipo de frases durante el proceso de extracción de la gramática, podríamos haber obtenido símbolos no terminales que no se encuentran directamente ligados a componentes lingüísticos reales, o incluso una gramática no independiente del contexto y, por tanto, no directamente tratable por nuestro analizador sintáctico.

Debido a esto, hemos preferido realizar una transformación más, la cual ha consistido en dividir el corpus SUSANNE en dos partes: una parte con todas las frases sin trazas (4.292 frases), a partir de la cual hemos extraído nuestra gramática, y otra parte con todas las frases con trazas (2.188 frases), la cual se ha utilizado como banco de experimentos.

Ejemplo 2.6 A continuación se muestra el aspecto general que presentan la parte de las frases sin trazas:

```
...
[ Fa [ CSf For CSf ] [ Ds:s [ DD1q each DD1q ] [ Po [ IO of IO ] [ Np [ DD2i
these DD2i ] [ NN2 lines NN2 ] Np ] Po ] Ds:s ] [ Vz [ VVZv meets VVZv ]
Vz ] [ YTL <bital> YTL ] [ N:o [ FOx Q FOx ] N:o ] [ YTR <eital> YTR ] [ P:p
[ II in II ] [ Np [ MC three MC ] [ NN2 points NN2 ] [ YC , YC ] [ REX
namely REX ] [ N@ [ MC two MC ] [ NN2 points NN2 ] [ P [ II on II ] [ FOx g
FOx ] P ] [ Ns+ [ CC and CC ] [ MC1 one MC1 ] [ NNL1n point NNL1n ] [ P [ II
on II ] [ Ms [ MC1 one MC1 ] [ Po [ IO of IO ] [ Np [ AT the AT ] [ JJ
multiple JJ ] [ NN2 secants NN2 ] Np ] Po ] Ms ] P ] Ns+ ] N@ ] Np ] P:p ]
Fa ]
...
```

y la parte de las frases con trazas:

```
...
[ Fa [ CSf For CSf ] [ Ni:s [ PPH1 it PPH1 ] Ni:s ] [ Vz [ VVZt includes
VVZt ] Vz ] [ Np:o173 [ AT the AT ] [ JJ emotional JJ ] [ NN2 ties NN2 ]
[ Fr [ CST that CST ] [ s173 [ YG - YG ] s173 ] [ V [ VV0v bind VV0v ] V ]
```

```
[ Np:o [ NN2 men NN2 ] Np:o ] [ P:q [ IIt to IIt ] [ Ns [ APPGh2 their
APPGh2 ] [ NN1c homeland NN1c ] Ns ] P:q ] Fr ] [ Np+:175 [ CC and CC ] [ AT
the AT ] [ JJ complex JJ ] [ NN2 motivations NN2 ] [ Fr [ CST that CST ]
[ s175 [ YG - YG ] s175 ] [ V [ VV0v hold VV0v ] V ] [ Ns:o [ AT1 a AT1 ]
[ JJ large JJ ] [ NNJ1c group NNJ1c ] [ Po [ IO of IO ] [ NNc people NNc ]
Po ] Ns:o ] [ R:w [ RL together RL ] R:w ] [ P:j [ IIa as IIa ] [ Ns [ AT1
a AT1 ] [ NN1c unit NN1c ] Ns ] P:j ] Fr ] Np+:175 ] Np:o173 ] Fa ]
...
```
□

Sobre la parte de frases sin trazas, hemos realizado todavía una transformación más, destinada a los siguientes propósitos. Primero, añadir un nuevo símbolo no terminal, `X`, que será el axioma o símbolo inicial de nuestra gramática. Segundo, sustituir las etiquetas de las palabras por su número correspondiente dentro del juego de etiquetas del corpus SUSANNE. Y finalmente, eliminar los símbolos no terminales que aparecen repetidos antes de cada corchete de cierre, con el fin de tener disponible el banco de árboles de referencia en el mismo formato de salida producido por nuestro analizador, y de esta manera facilitar la posterior comparación de los árboles de análisis.

Ejemplo 2.7 Este es, por tanto, el aspecto definitivo que presenta la parte de frases sin trazas, a partir de la cual se ha realizado la extracción de la gramática:

```
...
[ X [ Fa [ :33 For ] [ Ds_s [ :59 each ] [ Po [ :121 of ] [ Np [ :74 these ]
[ :172 lines ] ] ] ] [ Vz [ :403 meets ] ] [ :422 <bital> ] [ N_o [ :96
Q ] ] [ :423 <eital> ] [ P_p [ :106 in ] [ Np [ :146 three ] [ :172 points ]
[ :408 , ] [ :302 namely ] [ N@ [ :146 two ] [ :172 points ] [ P [ :106 on ]
[ :96 g ] ] [ Ns+ [ :14 and ] [ :147 one ] [ :182 point ] [ P [ :106 on ]
[ Ms [ :147 one ] [ Po [ :121 of ] [ Np [ :8 the ] [ :133 multiple ] [ :172
secants ] ] ] ] ] ] ] ] ] ]
...
```
□

La extracción de las reglas de la gramática a partir de este tipo de árboles parentizados se puede realizar mediante un sencillo algoritmo que hace uso de una estructura de pila inicialmente vacía, en la cual se van almacenando los símbolos que constituirán cada regla.

Algoritmo 2.1 El algoritmo de extracción de reglas gramaticales a partir de un árbol parentizado simplemente estudia los símbolos que forman parte de dicho árbol parentizado y considera los tres casos siguientes:

- Si aparece el símbolo `[` estamos ante el comienzo de una regla. Por tanto, se toma el siguiente símbolo del árbol parentizado, se añade a la parte derecha de la regla que está en la cima de la pila, y se crea un nuevo elemento en la pila para una nueva regla cuya parte izquierda es dicho símbolo.

- Si aparece el símbolo `]` estamos ante el final de una regla. Es decir, la regla que está en la cima de la pila en ese instante está ya completa, con lo cual se toma nota de ella y se elimina de la pila.

- Si aparece cualquier otro símbolo, necesariamente se tratará de una palabra, la cual simplemente se añade a la parte derecha de la regla que está en la cima de la pila en ese momento.

Los pasos básicos de este algoritmo son los siguientes:

```
function Extraer_Reglas (Árbol_Parentizado) =
  begin
    P ← Pila_Vacía;
    R ← Conjunto_Vacío;

    while (queden símbolos del Árbol_Parentizado por procesar) do
      begin
        A ← siguiente símbolo del Árbol_Parentizado;

        case A of
          [ :
            A ← siguiente símbolo del Árbol_Parentizado;
            Regla ← pop (P);
            Añadir A en la parte derecha de Regla;
            push (P, Regla);
            push (P, A → );

          ] :
            Regla ← pop (P);
            Añadir Regla a R;

          cualquier otro símbolo :
            Regla ← pop (P);
            Añadir A en la parte derecha de Regla;
            push (P, Regla)
        end
      end;

    return R
  end;
```

Este proceso de extracción genera gramáticas formadas por reglas *no parcialmente lexicalizadas*. En este tipo de gramáticas, los símbolos terminales aparecen sólo en reglas de la forma $A \rightarrow w_1\, w_2\, \ldots\, w_k$, donde A es una etiqueta y los w_i son símbolos terminales, es decir, palabras. La mayoría de las veces k será igual a 1. Los casos donde $k > 1$ corresponden a las unidades multipalabra, tales como palabras compuestas, expresiones hechas o locuciones. □

Ejemplo 2.8 La figura 2.2 es un ejemplo de la aplicación del algoritmo de extracción de reglas al árbol parentizado:

```
[ X [ S [ :89 There ] [ Vsb [ :368 was ] ] [ Ns_s [ :11 no ] [ :167
chance ] ] ] ]
```

y muestra que efectivamente las reglas gramaticales involucradas en él son:

```
X -> S                 :89 -> There
S -> :89 Vsb Ns_s      :368 -> was
Vsb -> :368            :11 -> no
Ns_s -> :11 :167       :167 -> chance
```

□

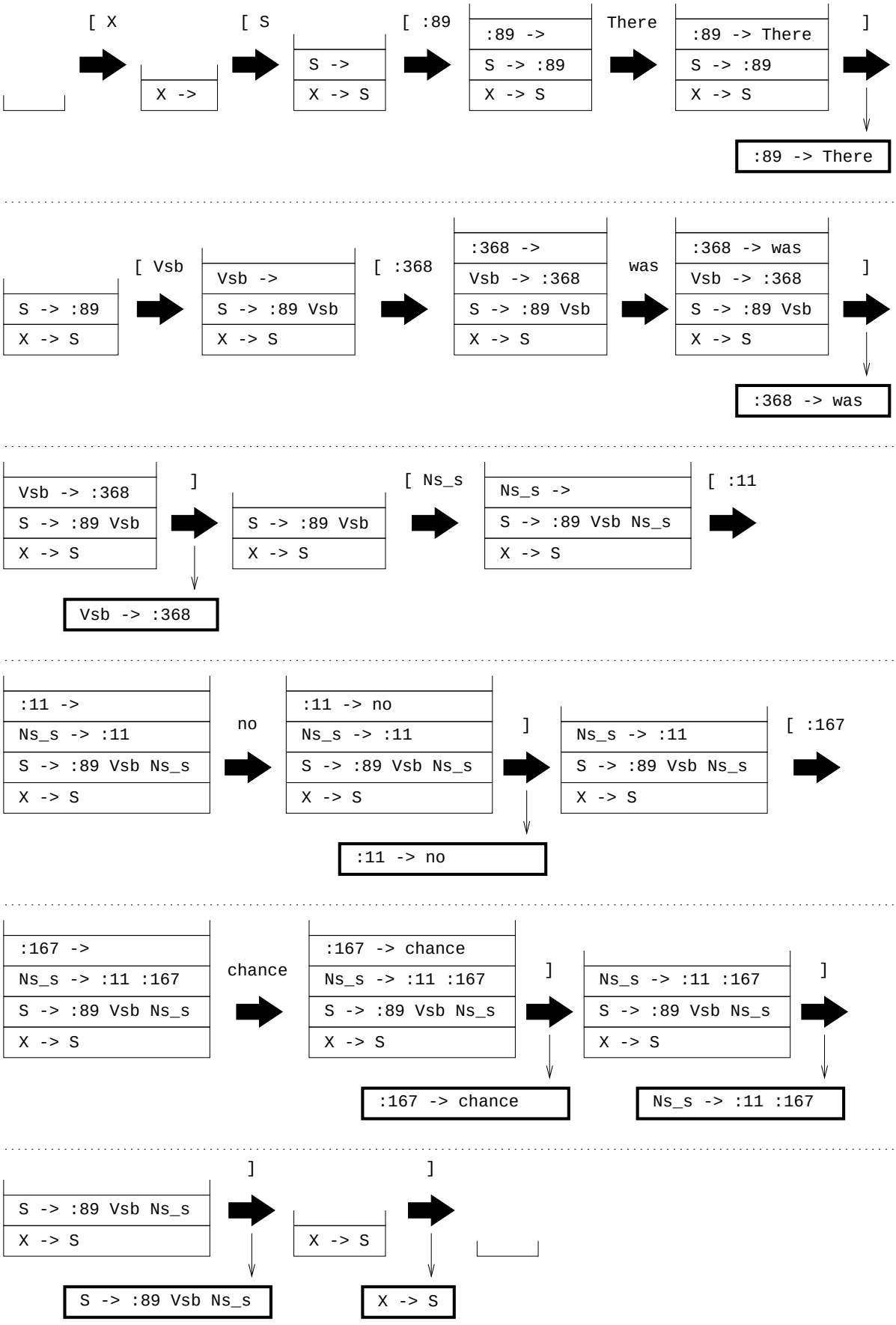

Figura 2.2: Ejemplo de ejecución de la función `Extraer_Reglas`

En la práctica, todas estas reglas se pueden entender como reglas gramaticales, pero no se suelen almacenar juntas. Generalmente, sólo forman parte de la gramática las reglas cuya parte izquierda es un símbolo no terminal, mientras que las reglas cuya parte izquierda es una etiqueta y por tanto cuya parte derecha es una palabra son precisamente las que constituyen el lexicón.

En todo caso, una vez que este proceso de extracción ha sido aplicado a todas las frases sin trazas, y tal y como indica el paradigma de análisis sintáctico estocástico que estudiaremos con detalle en el capítulo 8, se calcula la probabilidad de cada una de las reglas en relación al resto de producciones que comparten la misma parte izquierda. Es decir, se debe verificar que

$$\sum_{i=1}^{n} P(A \to \alpha_i) = 1$$

y, por tanto, la probabilidad de una regla genérica $A \to \alpha$ se calcula como:

$$P(A \to \alpha) = \frac{\text{número de veces que aparece la regla } A \to \alpha}{\sum_{i=1}^{n} \text{número de veces que aparece la regla } A \to \alpha_i}$$

donde: o bien A es cualquier símbolo no terminal, y entonces α y α_i son cualquier combinación de uno o varios símbolos no terminales y/o etiquetas; o bien A es una etiqueta, y entonces α y α_i son palabras. En ambos casos, n es el número de reglas diferentes cuya parte izquierda es A.

Antes de utilizar la gramática, hemos comprobado si contenía ciclos o no, ya que un ciclo podría hacer entrar al analizador sintáctico en un lazo sin fin. Esta comprobación se puede realizar mediante una sencilla operación de matrices. Dado que trabajaremos en todo momento con un analizador sintáctico ascendente, los ciclos sólo pueden ser producidos por las reglas unitarias, esto es, por reglas con un único símbolo no terminal tanto en la parte izquierda como en la parte derecha de la flecha. En el conjunto de reglas extraídas existían 206 reglas unitarias que involucraban a 187 símbolos no terminales. Se construyó entonces una matriz cuadrada U de $187 \times 187 = 34.969$ celdas, de tal manera que dados cualesquiera dos símbolos no terminales X e Y, la celda $U(X, Y)$ contenía la probabilidad de la regla $X \to Y$. Por tanto, 206 de esas celdas contenían las probabilidades de las reglas unitarias consideradas, mientras que el resto de las celdas estaban a cero. Sabiendo que

$$\sum_{i=0}^{\infty} x^i = 1 + x + x^2 + \ldots = \frac{1}{1-x}$$

calculamos entonces

$$V = \sum_{i=0}^{\infty} U^i = (I - U)^{-1} \tag{2.1}$$

donde I es la matriz identidad de orden 187. Si ahora se calcula $U \times V$, o lo que es lo mismo, $V - I$, se obtiene una matriz cuyas celdas contienen, para cualesquiera dos símbolos no terminales X e Y, la probabilidad de $X \overset{*}{\Rightarrow} Y$, es decir, la probabilidad de que X genere Y en un número finito de pasos. Por tanto, si existen valores diferentes de cero en la diagonal de $U \times V$, estamos ante la presencia de ciclos[10]. Después de la aplicación de este procedimiento, se detectó un único ciclo causado por las reglas `Ot` $\to$ `Nns` y `Nns` $\to$ `Ot`. La primera de estas reglas aparece numerosas veces a lo largo de las frases del corpus, mientras que la segunda aparecía una única vez en este punto concreto del corpus:

[10] Por supuesto, dado que este método de detección de ciclos está basado en series formales de endomorfismos, es decir, en correspondencias lineales, el cálculo propuesto no siempre es válido: la serie formal debe converger, es decir, $\lim_{i \to \infty} U^i = 0$. Esto es equivalente a decir que el radio espectral de la matriz U debe ser menor que 1 [Ciarlet 1988, Teorema 1.5.1, pp. 21-22]. La demostración de este teorema se encuentra también en [Ciarlet *et al.* 1991, Ejercicio 1.5.7, pp. 23-24]. En general, el *radio espectral* de una matriz A, que denotaremos

```
...
G22:0450j -AT The the [Nns[Ot[Np.
G22:0460a -JJ New new [Nns.
G22:0460b -NP1t York York .Nns]
G22:0460c -NNT2 Times time .Np]Ot]Nns]Po]
...
```

La forma de proceder fue suponer que esto se debía a un error de transcripción, reemplazar el subárbol `[ Nns [ Ot [ ... ] Ot ] Nns ]` por `[ Ot [ Nns [ ... ] Nns ] Ot ]` manualmente, e informar de tal anomalía a los constructores del corpus. Tras realizar de nuevo la extración de las reglas, este ciclo desapareció y obtuvimos una gramática directamente utilizable por nuestro analizador sintáctico. Esta gramática está compuesta por 17.669 reglas y 1.525 símbolos no terminales, y la representación arborescente de las reglas (veáse la sección 8.4.7) contiene 28.117 nodos.

Ejemplo 2.9 A continuación, para mostrar el aspecto general que presentan las reglas, incluimos un pequeño conjunto de líneas tomadas directamente de la gramática:

```
X -> Fa (0.002796)                    X -> Fa% (0.000233)
X -> Fa_121 (0.000233)                X -> Fc (0.000466)
X -> Ff (0.000233)                    X -> Fr (0.000233)
X -> L (0.009320)                     X -> L! (0.000233)
X -> L+ (0.000233)                    X -> L? (0.000233)
X -> L?+ (0.000466)                   X -> S (0.916356)
X -> S! (0.001165)                    X -> S!+ (0.000233)
X -> S% (0.000233)                    X -> S* (0.011650)
X -> S*+ (0.000699)                   X -> S+ (0.029590)
X -> S? (0.017008)                    X -> S?+ (0.000932)
X -> S@ (0.000699)                    X -> S_129 (0.000233)
X -> S_133 (0.000233)                 X -> S_145 (0.000233)
X -> S_149 (0.000233)                 X -> S_151 (0.000233)
X -> S_205 (0.000233)                 X -> S_209 (0.000233)
X -> S_223 (0.000233)                 X -> Tb (0.000466)
X -> Tb! (0.000233)                   X -> Tb? (0.000466)
X -> Tg (0.002563)                    X -> Tg+ (0.000233)
X -> Ti (0.000466)                    X -> Tn (0.000466)
A -> :27 Ni_s Vzb P_r (0.105263)      A -> :27 Ns_S Vzp (0.0526316)
A -> :27 P_p (0.0526316)              A -> :27 R_t (0.0526316)
A -> :27 R_t Jh_e (0.0526316)         A -> :27 Ti_z (0.0526316)
```

mediante $\rho(A)$, se calcula como sigue. Dada A, una matriz cuadrada de orden n, calculamos el determinante $|A - \lambda I|$, donde I es la matriz identidad también de orden n. Lo que obtenemos no es un valor concreto, sino un polinomio de grado n en λ, de la forma $a_n\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0$. Dicho polinomio se denomina *polinomio característico* de A, y sus raíces, $r_1, r_2, \ldots, r_n$, son los *valores propios* de A. Pues bien, si de cada uno de los valores propios consideramos su valor absoluto, el radio espectral de la matriz A será el máximo de esos valores absolutos. Es decir, $\rho(A) = \max\{|r_i|, \text{tal que } r_i, 1 \leq i \leq n, \text{es un valor propio de } A\}$. Por tanto, si $\rho(U) < 1$, el sumatorio de la ecuación (2.1) tiene sentido y converge efectivamente hacia el valor $(I - U)^{-1}$. Pero en definitiva, es suficiente con intentar el cálculo de la inversa de la matriz $I - U$. Si dicha matriz no resulta inversible, la matriz V debe ser calculada a través de un procedimiento iterativo que acumule la suma de las sucesivas potencias de U, tal y como se deduce de los métodos de búsqueda de caminos propuestos en la teoría de grafos [Grassmann y Tremblay 1996, Teorema 7.4, pp. 365-366], aunque obviamente no consideraremos las potencias hasta infinito, sino hasta el número que nos interese para volver al punto de partida y detectar así los ciclos (en nuestro caso, hasta la potencia 187).

```
A -> :27 Vg Fn_o (0.0526316)          A -> :27 Vn P_p (0.105263)
A -> :27 Vn P_q (0.0526316)           A -> :27 Vn P_r (0.0526316)
A -> :27 Vn P_u (0.0526316)           A -> :27 Vn Pb_a (0.315789)
A+ -> :14 Vn R_q P_q (0.5)            A+ -> :20 :27 Np_s V Ns_o (0.5)
A_c -> :27 Ni_s Vd R_n (0.333333)     A_c -> :27 Ni_s Vsb (0.333333)
A_c -> :27 Ns_s Vd Ni_o (0.333333)    ...                                  □
```

2.3.2.2 Dos *corpora* etiquetados

Para ser utilizado como recurso de referencia por las herramientas de etiquetación ya existentes, es también sencillo obtener un corpus etiquetado a la manera tradicional a partir del corpus SUSANNE, fijándonos únicamente en los nodos hoja, es decir, en las palabras y en sus etiquetas correspondientes, y despreciando la información relativa al resto de nodos. De esta forma, tendremos un buen marco de pruebas para el estudio comparativo de las técnicas de etiquetación tradicionales y las propuestas en este trabajo. Siguiendo la misma filosofía utilizada a la hora de extraer la gramática, hemos generado dos *corpora* etiquetados: uno correspondiente a la parte de frases sin trazas y otro a la parte de frases con trazas. A continuación se describen las características de cada uno de ellos:

- El corpus etiquetado correspondiente a la parte de frases sin trazas tiene 4.292 frases y 77.275 palabras, es decir, un número medio de 18 palabras por frase. El tamaño del fichero es de 751.359 caracteres. Las características del lexicón o diccionario constituido por todas las formas que aparecen en este corpus son las siguientes:

10.938 formas con 1 etiqueta,	864 formas con 2 etiquetas,
87 formas con 3 etiquetas,	27 formas con 4 etiquetas,
9 formas con 5 etiquetas,	5 formas con 6 etiquetas,
2 formas con 7 etiquetas,	1 forma con 8 etiquetas y
2 formas con 9 etiquetas.	

  Esto es, 11.935 formas diferentes, con 13.150 etiquetas posibles. Si calculamos el porcentaje de formas ambiguas y el número medio de etiquetas por forma, obtenemos el siguiente resultado:

  $$\%\ \text{formas ambiguas} = \frac{\#\ \text{formas ambiguas}}{\#\ \text{formas}} \times 100 = \frac{11.935 - 10.938}{11.935} \times 100 = 8,35\ \%$$

  $$\#\ \text{medio de etiquetas por forma} = \frac{\#\ \text{etiquetas}}{\#\ \text{formas}} = \frac{13.150}{11.935} = 1,10\ \text{etiquetas por forma.}$$

  Mucho más interesante es calcular las mismas características directamente con todas las palabras del corpus, y obtenemos las siguientes cifras:

42.550 palabras con 1 etiqueta,	13.004 palabras con 2 etiquetas,
9.225 palabras con 3 etiquetas,	2.175 palabras con 4 etiquetas,
2.174 palabras con 5 etiquetas,	2.015 palabras con 6 etiquetas,
1.539 palabras con 7 etiquetas,	2.574 palabras con 8 etiquetas y
2.019 palabras con 9 etiquetas.	

  Esto es, 77.275 palabras, con 177.429 etiquetas posibles. Si calculamos de nuevo el porcentaje de palabras ambiguas y el número medio de etiquetas por palabra, obtenemos entonces:

$$\%\text{ palabras ambiguas} = \frac{\#\text{ palabras ambiguas}}{\#\text{ palabras}} \times 100 = \frac{77.275 - 42.550}{77.275} \times 100 = 44,93\ \%$$

$$\#\text{ medio de etiquetas por palabra} = \frac{\#\text{ etiquetas}}{\#\text{ palabras}} = \frac{177.429}{77.275} = 2,30 \text{ etiquetas por palabra.}$$

- El corpus etiquetado correspondiente a la parte de frases con trazas tiene 2.188 frases y 60.759 palabras, es decir, un número medio de 28 palabras por frase. El tamaño del fichero es de 593.735 caracteres. Las características del lexicón o diccionario constituido por todas las formas que aparecen en este corpus son las siguientes:

  9.322 formas con 1 etiqueta, 687 formas con 2 etiquetas,
  70 formas con 3 etiquetas, 21 formas con 4 etiquetas,
  8 formas con 5 etiquetas, 3 formas con 6 etiquetas,
  2 formas con 7 etiquetas, 1 forma con 8 etiquetas y
  1 formas con 14 etiquetas.

  Esto es, 10.115 formas diferentes, con 11.084 etiquetas posibles. Si calculamos el porcentaje de formas ambiguas y el número medio de etiquetas por forma, obtenemos el siguiente resultado:

$$\%\text{ formas ambiguas} = \frac{\#\text{ formas ambiguas}}{\#\text{ formas}} \times 100 = \frac{10.115 - 9.322}{10.115} \times 100 = 7,84\ \%$$

$$\#\text{ medio de etiquetas por forma} = \frac{\#\text{ etiquetas}}{\#\text{ formas}} = \frac{11.084}{10.115} = 1,06 \text{ etiquetas por forma.}$$

  Mucho más interesante es calcular las mismas características directamente con todas las palabras del corpus, y obtenemos las siguientes cifras:

  34.329 palabras con 1 etiqueta, 8.957 palabras con 2 etiquetas,
  3.580 palabras con 3 etiquetas, 2.232 palabras con 4 etiquetas,
  4.984 palabras con 5 etiquetas, 2.268 palabras con 6 etiquetas,
  2.820 palabras con 7 etiquetas, 1.236 palabras con 8 etiquetas y
  353 palabras con 14 etiquetas.

  Esto es, 60.759 palabras, con 145.009 etiquetas posibles. Si calculamos de nuevo el porcentaje de palabras ambiguas y el número medio de etiquetas por palabra, obtenemos entonces:

$$\%\text{ palabras ambiguas} = \frac{\#\text{ palabras ambiguas}}{\#\text{ palabras}} \times 100 = \frac{60.759 - 34.329}{60.759} \times 100 = 43,50\ \%$$

$$\#\text{ medio de etiquetas por palabra} = \frac{\#\text{ etiquetas}}{\#\text{ palabras}} = \frac{145.009}{60.759} = 2,39 \text{ etiquetas por palabra.}$$

La figura 2.3 resume gráficamente el proceso que hemos seguido para extraer los recursos lingüísticos a partir del corpus SUSANNE, junto con las características básicas de los mismos.

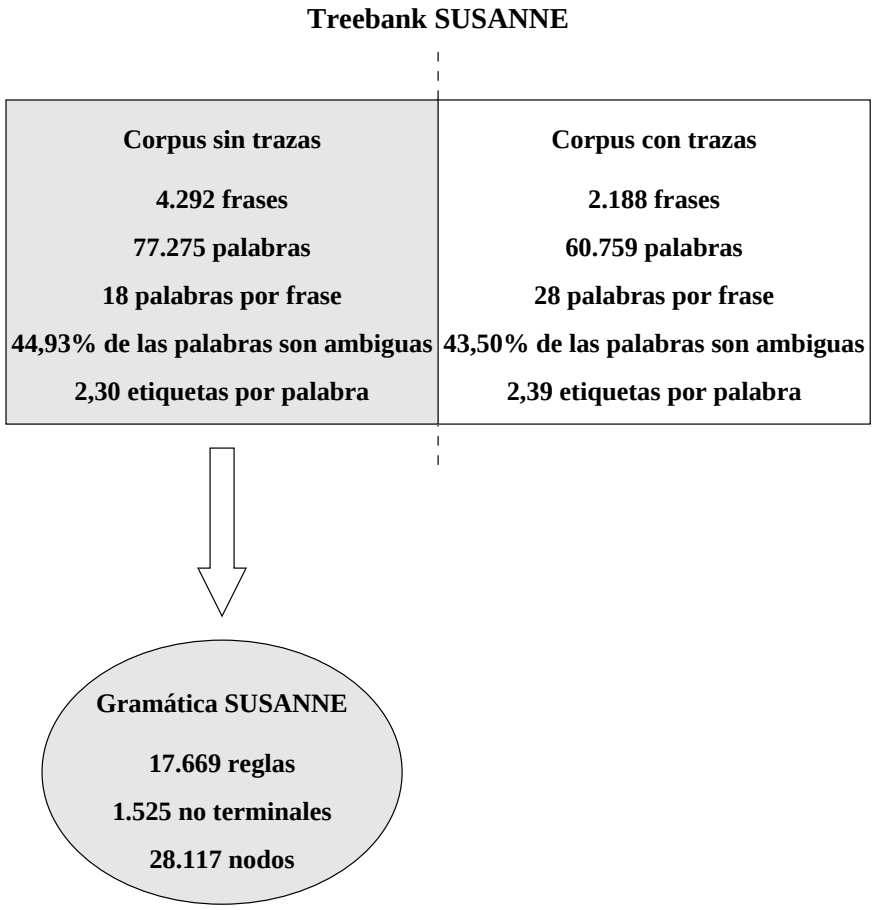

Figura 2.3: División y extracción de recursos lingüísticos a partir del corpus SUSANNE

Los siguientes capítulos explicarán detalladamente cómo hemos utilizado todos los recursos lingüísticos presentados aquí, no sólo para testear la calidad de los mismos, sino también para evaluar el rendimiento de las distintas técnicas de etiquetación, tanto las tradicionales como las de nuevo diseño.