INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA
UNIVERSITAT POMPEU FABRA

Programa de doctorado: Lingüística Aplicada
Bienio 2000 - 2002

Tesis doctoral

Los sintagmas nominales extensos especializados en inglés y en español: Descripción y clasificación en un corpus de genoma

Gabriel Ángel Quiroz Herrera

Tesis doctoral
Para optar al título de doctor por la Universitat Pompeu Fabra

Dirigida por: Mercè Lorente Casafont

UNIVERSITAT POMPEU FABRA

Barcelona, 2008

Dipòsit legal: B.40585-2008
ISBN: 978-84-692-0978-3

Agradecimientos

Las tesis suelen ser un trabajo individual, pero que sin la ayuda de otros colegas e instituciones es difícil que llegue a buen puerto.

Quiero agradecer muy especialmente a mi tutora Mercè Lorente por aceptar el tema; ¡al final hemos pasado de 14 a 7 tokens! Debo agradecer las largas tutorías, su punto de vista siempre tan crítico y las discusiones académicas en aspectos lingüísticos que han nutrido esta tesis y también por darme su apoyo en momentos críticos.

A Teresa Cabré por aceptarme en el IULA durante varios años y hacerme partícipe de muchas actividades del Instituto.

A Berta Nelly Cardona, por toda su colaboración para venir a Barcelona y a Michael Scholand y Lidia Cámara de Star SL por permitirme estar en su empresa durante estos años.

A Jorge Vivaldi por ayudarme en la extracción y preparación de datos *ad libitum*. Sin su ayuda y paciencia la fase de datos no hubiera salido bien.

A la Real Academia Española por permitirme acceder a los datos del CREA y al señor Fernando Sánchez por facilitarme la consulta de los patrones en español.

A Ricardo Guantiva, Pedro Patiño y Diego Burgos por las revisiones y críticas hechas al manuscrito final; todos los gazapos son, por supuesto, míos.

A mis profesores del *Institut Universitari de Lingüística Aplicada*-IULA, Teresa Cabré, M. Teresa Turell, Carmè Bach, Lluís De Yzaguirre, Rosa Estopà y

Jaume Llopis por los aportes directos e indirectos dados a esta tesis y a los profesores de otros centros quienes han criticado versiones previas de este trabajo o me han aportado material bibliográfico: Ulrike Oster, Enilde Faultisch, Ana María Cardero, Isabel Karely, Natividad Gallardo, Melva Márquez, Elena Bogomilova, Cleci Bevilacqua y Hada Salazar.

Finalmente, a mis colegas y amigos Carlos Muñoz, John Jairo Giraldo, Carles Tebé, Diego Burgos, Juan Manuel Pérez, Pedro Patiño, Ricardo Guantiva, Roxana Folguerà, Araceli Alonso, Ana Corrales, Fernando Yuste, Beatriz Pineda, Anderson Franco y otros tantos que se encuentran al otro lado del Atlántico, con quienes he compartido largas conversaciones sobre mi tesis y otros temas, no menos interesantes.

Mónica, gracias por estar a mi lado durante estos años y superar todas las vicisitudes que uno puede tener fuera de su país, por el amor y el cariño y los buenos momentos que hemos disfrutado en este país.

A mis hijos Miguelangel y Felipe,
por adaptarse tan bien a las culturas catalana y española
sin olvidar nuestras culturas, la antioqueña y la colombiana.

A mi padre Gabriel y mi madre Beatriz,
y a mis hermanos Adriana, Walter, Edison y Víctor
por el todo cariño que nos han brindado en la lejanía.

Migue y Pipe, sólo me faltaba por escribir
esta página para terminar la tesis.

Abreviaturas

Adj.: adjetivo
Adv.: adverbio
Conj.: conjunción
D: determinante
N: nombre, sustantivo
PP: participio pasado
PPi: participio presente
Prep.: preposición
V: verbo
SNEE: sintagma nominal extenso especializado
SA: sintagma adjetival
SN: sintagma nominal
SP: sintagma preposicional
UMLS: Unified Medical Language System
UMLSKS: Unified Medical Language System Knowledge Source Server
POS: Part of speech
IULA: Institut Universitari de Lingüística Aplicada
DRAE: Diccionario de Real Academia Española
WN: WordNet
EWN: EuroWordNet
IMMRAD: Introduction, Materials and Methods, Results and Discussion
ESP: español
ENG: inglés
SN: sintagma nominal
SAdj.: sintagma adjetival
SPrep.: sintagma preposicional
IFCC: Diccionario inglés-español de Ciencias de Laboratorio Clínico de la Federación Internacional de Química Clínica
IMF: International Monetary Fund
RD: Routledge Spanish Dictionary of Business, Commerce and Finance
ISI: International Statistical Institute Multilingual Glossary of Statistical Terms

RAE: Real Academia Española

CREA: Corpus de Referencia del Español Actual

POS: Part of speech

Porc.: porcentaje

Índice de contenidos

1. Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma

1.1 Introducción

En la presente tesis doctoral se describen y clasifican los sintagmas nominales extensos especializados (SNEE) de más de tres *tokens* en inglés y en español en textos especializados del nivel experto-experto en el área del genoma.

En esta tesis se quiere corroborar los resultados preliminares obtenidos en trabajos anteriores para observar su comportamiento y así llegar a establecer unas regularidades, de modo que los profesionales de las lenguas inglesa y española puedan tener una herramienta que les permita solucionar adecuadamente los sintagmas nominales con premodificación extensa.

En resumen, los fenómenos que se analizarán en esta tesis pueden resumirse básicamente en: a) el estudio de la premodificación compleja de los sintagmas nominales en inglés; b) el estudio de la posmodificación en español; c) el análisis de la correlación entre las estructuras y las categorías semánticas; d) las tendencias semánticas de estos sintagmas en cuantos a clases y patrones; e) el comportamiento de los sintagmas nominales extensos en diccionarios especializados y corpus de referencia; f) y, finalmente las tendencias de traducción de este tipo de sintagmas en un corpus paralelo.

1.2 Antecedentes de la tesis

Esta tesis es el fruto de diversos trabajos realizados en el marco del doctorado en Lingüística Aplicada del *Institut Universitari de la Universitat Pompeu Fabra*, bajo la supervisión de la Prof[a]. Dra. Mercè Lorente.

La idea inicial de esta tesis surgió como un problema de traducción en la vida profesional al observar que en las traducciones científico-técnicas aparecían con cierta regularidad sintagmas nominales con premodificación compleja. Además, su traducción se convertía en un problema por su complejidad sintáctico-semántica, por la falta de fuentes de consulta para resolverlos y por el poco tratamiento dado en la literatura de traducción hasta ese entonces.

Posteriormente, durante mis estudios de postgrado en la Escuela de Idiomas de la Universidad de Antioquia-Colombia y luego como profesor de la asignatura de Traducción Científico-técnica en el programa de Traducción de la misma universidad encontré que muchos otros colegas compartían la misma dificultad y que también carecían de herramientas para enseñar a resolverlos con propiedad a los estudiantes. Con los años fui coleccionando docenas de ellos como piezas de museo.

En aquel entonces, influido por las corrientes funcionalistas de la traducción (Hölz-Mantari 1984, Nord 1991), pretendía que los estudiantes identificaran las características lingüísticas de los diferentes géneros textuales que componían la asignatura. Las referencias que se tenían eran pocas y las soluciones eran muy intuitivas. Llegué a la conclusión de que existía muy poca descripción lingüística de este tipo de géneros del inglés al español. Incluso hoy en día la descripción sigue siendo poca si se compara con lo realizado en otras lenguas. Ya en esa época, empezaba a tomar fuerza en español una corriente teórica que defendía la idea de la traducción como disciplina autónoma independiente de la lingüística y daba por sentado que el traductor debía tener la competencia lingüística al momento de traducir. Por tanto, se cerraba la posibilidad de hacer investigación de características lingüísticas para propósitos de traducción. Por otro lado, fruto de todas mis lecturas y el contacto con colegas europeos y americanos, descubrí que muchos de ellos no provenían de la traducción como tal por diversas razones. Igualmente, descubrí que los aportes más interesantes de los últimos años, los habían hecho autores que tenían una formación en lingüística y sus disciplinas colindantes, la sociología, la

psicología, la informática, entre otras. Por tanto, una explicación a este fenómeno que pudiera satisfacer parte del problema tanto para aspectos investigativos, profesionales y didácticos debería ser lingüística.

En el doctorado, comencé a indagar la manera de abordar el problema. Dada la estructura del doctorado de mi bienio, se realizaron varios trabajos de aproximación al problema y el estudio piloto que me permitiera evaluar las posibilidades del tema de tesis y las dificultades potenciales a las que podría enfrentar.

En un primer trabajo, se exploró la extensión de los patrones en inglés, la categoría predominante en la premodificación y patrones más frecuentes. En este primer acercamiento se vio la necesidad de ampliar el corpus para poder constatar si las tendencias continúan pues la extensión promedio de la premodificación y la categoría premodificadora así como algunos patrones no coincidían con lo encontrado por otros autores (Biber *et al* 1999, Montero 1995).

En un segundo trabajo (Quiroz *et al* 2004), se exploró la posibilidad de encontrar regularidades de traducción en un minicorpus. A pesar del tamaño del corpus, se pudo observar las posibilidades que presenta un patrón en inglés y sus soluciones en español en algunos patrones y así, sacar el máximo de regularidades en un corpus de mayor tamaño.

En el proyecto presentamos fundamentalmente la metodología que se llevaría a cabo en la tesis doctoral. Fruto de este proyecto de tesis y los comentarios y sugerencias del tribunal, se refinó la metodología presentada aquí. De igual modo, algunas partes de esta tesis ya han sido publicadas (Quiroz 2005 y Quiroz 2006) y se han ampliado aquí y, por tanto, haremos referencia a ellas.

1.3 Objeto de análisis

Newmark (1988: 39-41) hace una clasificación de los tipos de texto de acuerdo con las funciones del lenguaje de Jakobson (1959). Una de ellas es la función informativa, propia de los textos científico-técnicos. Dentro de este tipo de textos, Newmark establece 3 niveles de formalidad de acuerdo con la relación emisor-destinatario. En los tres niveles descritos, experto-lego, experto-técnico y experto-experto, define una serie de características lingüísticas diferenciadoras de su función pragmática. En el nivel experto-experto, una de las características que más resalta Newmark *a priori* es la que denomina "multi-noun compounds".

Otros autores también ven este fenómeno como una característica lingüística relevante del discurso especializado o la traducción científico-técnica (Sager *et al.* 1980; Horsella y Pérez 1990; Vivanco 1996; Quiroz y Muñoz 1997; Abril y Ortiz 1998; Cartagena 1998; Linder 2002; Gotti 2003 y Scarpa 2001).

Para restringir nuestro objeto de estudio, definiremos un sintagma nominal extenso especializado[1] de la lengua inglesa o española como una frase nominal definida o indefinida de 3 o más tokens (2 premodificadores o posmodificadores y un núcleo). Un SNEE consta de un sustantivo nuclear (núcleo) precedido por diversos elementos (premodificación) en inglés, en especial por determinantes, adjetivos, participios de pasado y presente, sustantivos, en algunos casos, por adverbios, u otras categorías y otros elementos no verbales como símbolos, signos, etc. y seguido en español (posmodificación), en especial por adjetivos, preposiciones, sustantivos, etc. Estas unidades suelen ser términos o tener términos o conjuntos de términos en su estructura combinados, en algunos casos, con unidades del lenguaje común.

[1] En esta tesis, nos referiremos a los sintagmas nominales extensos especializados también como sintagmas nominales extensos o sintagmas nominales.

Se ha establecido para este estudio un número mínimo de 3 tokens (2 premodificadores o posmodificadores mínimo) ya que los estudios realizados en corpus desde una perspectiva de traducción o terminología no presentan datos sobre sintagmas de más de 2 ó 3 elementos premodificadores o posmodificadores. Por otro lado, los corpus de estos estudios no incluyen un porcentaje significativo del tipo de texto que se trata en esta tesis. Por tanto, se ha analizado desde 3 tokens para observar las características sintáctico-semánticas y la frecuencia de estos sintagmas.

Para la descripción del objeto de estudio, se mantendrá la postura de que no existe una dicotomía entre lenguaje general y "lenguaje" especializado (o mejor dicho, discurso especializado); no es un problema de "otro" lenguaje con características excluyentes o con una gramática especial sino que se trata simplemente de un subsistema de la lengua.

Los diferentes análisis se llevarán a cabo desde la óptica de una gramática general y sólo su uso y frecuencia en el discurso especializado se debe a razones pragmáticas, conceptuales o cognitivas como algunos autores también lo afirman (Cartagena 1998; Cabré 1999; Gotti 2003; Scarpa 2001).

Para ilustrar el objeto de estudio se han coleccionado los siguientes ejemplos (1 a 4) tomados de las diferentes gramáticas (de referencia) y libros de sintaxis[2]:

1. the beautiful big old neglected square red Jamaican stone plantation houses (11 palabras, 9 en posición premodificadora*) (Let's Write English* 1980)

2. an attractive tight-fitting brand-new pink Italian lycra women's swimsuit (11 palabras, 10 en posición premodificadora) (*The Cambridge Grammar of the English Language* 2002: 543)

[2] Estos ejemplos se han empleado en otros trabajos, pero nos parece oportuno presentarlos en esta tesis.

3. horseback riding school cafeteria breakfast menu substitution list (Levi 1978: 5)

4. the hydrophobic polyethylene terephthalate (PET) and hydrophilic (water-soluble) polyethylene (PEO) blocks (Sager, Dungworth, and McDonald 1980: 221)

Los ejemplos de 5 a 12 se han coleccionado de textos de diferentes áreas temáticas.

5. l-(-)2',3'-Dideoxythiacytidine (l(-)SddC, Lamivudine) resistant hepatitis B virus (HBV). (10 tokens, 7 en posición premodificadora).
6. *Autographa californica* nuclear polyhedrosis virus (AcMNPV)-infected cells (8 palabras, 7 en posición premodificadora)
7. a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model (13 palabras, 11 en posición premodificadora)
8. an aperture emission mode scanning near-field optical microscope (SNOM)
9. a 30-day-no-questions asked product return policy
10. the Thermo-Sequenase fluorescent-labeled primer cycle sequencing kit
11. a human acute lymphoblastic leukemia CCRF-CEM cDNA library
12. *in vitro*-translated, radioactively labeled wild type and mutant Sox10 proteins

Los ejemplos de 13 a 20 se tomaron del corpus de genoma del IULA - *Institut Universitari de Lingüística Aplicada* de la *Universitat Pompeu Fabra*.

13. A novel, red, low-potential, periplasmic copper protein
14. a membrane bound proton-translocating pyrophosphatase
15. a green fluorescent protein (GFP) fusion protein
16. bright clear small dot-like fluorescent structures
17. the trans-Golgi localized reversibly glycosylated polypeptide (RGP1)
18. an efficient nuclear magnetic resonance (NMR)-based metabonomic approach
19. *Drosophila melanogaster* maternal nuclear protein kinase Dm-nk
20. The mitochondrial inner membrane AAA metalloprotease family

En español también podemos encontrar sintagmas nominales de esta longitud que pueden ser términos, es decir, unidades especializadas. Por supuesto que no hay una premodificación como en inglés dadas las características del español, pero que nos muestra que el concepto de sintagma nominal especializado extenso va más allá de los límites considerados "normales". A guisa de ejemplo, se presentan algunos sintagmas nominales especializados extensos del Diccionario Espasa de Medicina (1999: 320).

Como entrada principal se encuentran los ejemplos 21 a 25.

21. déficit familiar de lecitín-colesterol-acil-transferasa
22. déficit neurológico isquémico reversible
23. déficit selectivo de subclases de IgG
24. déficit de alfa-1-antitripsina
25. años de vida ajustados según la calidad

Como subentrada se encuentran los ejemplos 26 a 28.

26. ácido graso Omega 3
27. acidosis tubular renal distal hiperpotasémica
28. rigidez muscular arteriosclerótica de Forester

En ámbitos técnicos también se pueden encontrar sintagmas extensos especializados como los ejemplos tomados de los diccionarios en CD-ROM *Spanish Dictionary of Business, Commerce and Finance* - Diccionario Inglés de Negocios, Comercio y Finanzas (1998) y *Spanish Technical Dictionary* (1998) ambos de Routdledge, Mosby inglés español (2000) y *The IEC Multilingual Dictionary* (2005), como los ejemplos de 29 a 40:

29. adaptive differential pulse coded modulation (modulación adaptable diferencial de impulsos en código)
30. ammonium nitrate fuel oil (explosivo compuesto de nitrato amónico y fueloil)
31. broadband-integrated services digital network (red digital de servicios integrados de banda ancha)

32. compact disc programmable read-only memory (disco compacto de memoria de sólo lectura programable)
33. containerized lighter aboard ship system (sistema de buques remolcadores para transporte de gabarras cargadas con contenedores)
34. consolidated link-layer management message (mensaje de gestión consolidada de enlace entre capas)
35. extended binary-coded decimal-interchange code (código ampliado de caracteres decimales codificados en binario)
36. permanent income bearing share (acción productora de renta permanente)
37. directly unproductive profit-seeking activities (actividades directamente improductivas con fines lucrativos)
38. adult respiratory distress syndrome [ARDS] (síndrome de dificultad respiratoria del adulto [SDRA])
39. contoured adducted trochanteric controlled alignment method [CAT-CAM] (método de alineación controlada del contorno trocantérico en aducción [CAT-CAM])
40. quadrature-axis sub-transient open-circuit time constant (constante de tiempo subtransitoria transversal en circuito abierto)

En otras lenguas como el francés y el alemán se puede dar cuenta también de este fenómeno. Por ejemplo, Kocourek (1991: 140) cita el fenómeno, tomando ejemplos de otros autores como se presenta de 41 a 44.

41. maillot de bain féminin d'une seule pièce dégageant les côtes, les bas du dos et les hanches (Rey-Devobe 1973: 92)
42. Donaudampfschiffahrtsgesellchaftskapitänwitwenrentenauszahlungstag (Heger 1971: 72)
43. VDI Ultrakurzwellenüberreichweitenfernfunkverbindung (Heger 1971: 72)
44. valve spool dirt excluding rubber washer steel spacer (Horecký 1963: 275)

1.4 Hipótesis y supuestos de partida

En esta tesis se parte de la idea de que la noción de sintagma nominal extenso especializado es el resultado de una serie de parámetros y rasgos que afectan:

- Las características de cada uno de los elementos del sintagma nominal extenso especializado
- Las relaciones que se establecen entre los diferentes elementos que constituyen un sintagma nominal extenso especializado y éste percibido como una unidad
- La relación entre el sintagma nominal extenso especializado y el resto de la oración
- La función del sintagma nominal extenso especializado dentro de la oración.

Además, la noción de sintagma nominal extenso especializado presenta rasgos diferenciales respecto de la noción general de SN:

- La extensión de los sintagmas nominales extensos especializados es mayor en los ámbitos científico-técnicos.

- Los SNEE presentan una diferencia estadística mayor en determinados registros y tipos de texto que lo propuesto por la bibliografía.

Las condiciones de interacción entre los interlocutores, reflejadas en el tipo de texto, hacen de los sintagmas nominales extensos especializados una característica muy relevante del discurso científico como lo proponen Swales (1974: 129) y Halliday (1998: 193).

Los SNEE son un elemento fundamental en la cohesión y compactación del discurso especializado como lo propone Ormrod (2001: 9-23).

Hasta donde se ha explorado en trabajos previos (Quiroz, Lorente, Yzaguirre 2004; Quiroz 2005a, 2005b y 2006), los SNEE se manifiestan a través de una variedad de estructuras no descritas sistemáticamente para el par inglés-español.

Finalmente, los SNEE pueden detectarse y extraerse de los textos a partir de criterios léxico-sintácticos y semánticos.

De acuerdo con Cabré (2003: 46-50), un texto especializado tiene unas condiciones en la estructura textual como la estructura informativa que tiene a su vez, unas características graduables como la precisión, la concisión, la sistematicidad, la objetividad y la impersonalidad que hacen que un texto sea más o menos especializado. Los sintagmas nominales extensos especializados pueden vehicular una gran densidad y una alta precisión en un texto de especialidad mediante los elementos internos de un sintagma nominal (acrónimos, símbolos, códigos, clasificaciones, compactación conceptual de la premodificación, entre otros).

Para esta tesis se proponen las siguientes hipótesis:

1. Los sintagmas nominales extensos especializados no son un problema del discurso especializado, son un fenómeno de la lengua que presenta mayor frecuencia en el discurso especializado y que tiene unas características sintáctico-semánticas determinadas.

Como puede verse en el estado de la cuestión §3, muchos autores ven este fenómeno como un problema, una aberración de la lengua o una falta de estilo básicamente desde una perspectiva prescriptivista. Dentro del marco de la teoría comunicativa de la terminología (Cabré 1999: 34-35) y en consonancia

con los principios metodológicos f y g[3], esta tesis se enmarca dentro una perspectiva descriptivista en la cual las unidades se extraen de textos reales que reflejan las condiciones comunicativas de un grupo de hablantes expertos o de un área en un nivel de especialidad determinado como parte de la lengua general. Por tanto, se reconoce que estas unidades son un fenómeno natural dentro de la situación comunicativa en la que se circunscribe y es necesario buscar las regularidades en su comportamiento y sistematizar su interpretación.

2. Los sintagmas nominales extensos especializados pueden describirse, clasificarse, explicarse y predecirse desde la gramática de una lengua como todos los fenómenos lingüísticos de los discursos de los ámbitos de especialidad.

Partimos de la idea planteada por Cabré (1993: 177) al hablar de los sintagmas terminológicos y libres en cuanto a que todos los fenómenos lingüísticos de los discursos especializados pueden explicarse mediante el uso de la gramática de la lengua general. La dicotomía, ya clásica, entre lenguaje general y lenguaje especializado o entre gramática de la lengua general y la gramática de los "lenguajes especializados" no se tendrá en cuenta en esta tesis puesto que la diferencia radica en que hay áreas de especialidad que usan determinados recursos lingüísticos, como los sintagmas nominales extensos especializados, con mayor frecuencia debido a razones pragmáticas de la

[3] f) El método es necesariamente descriptivo y consiste en la recopilación de las unidades reales usadas por los especialistas de un campo en distintas situaciones de comunicación. Esta diversidad de situaciones presupone que el corpus de extracción de los términos debe ser heterogéneo y representativo. Ello no impide que para un trabajo determinado pueda ser homogéneo tanto en su nivel de especialización y en el tipo de textos seleccionados, como también en la perspectiva de tratamiento del tema. Los términos seleccionados son unidades reales, no necesariamente satisfactorias ni normalizadas, simplemente reales. Solo en caso de que el trabajo pretenda ser una representación de los términos normalizados, se entrará en la actividad de fijación de una variante y/o reducción de la variación denominativa.

g) Las unidades retenidas en los textos como representativas del conocimiento especializado pueden ser términos (nominales, verbales, adjetivales) o unidades más amplias, combinaciones frecuentes en un determinado ámbito de especialidad. Estas combinaciones pueden ser terminológicas (unidades polilexemáticas denominativas de un concepto), fraseológicas (sintagmas no autónomos comunicativamente, que contienen por lo menos un término, habitualmente verbales frecuente y específicamente usados en una materia), o combinaciones aun más vastas que constituyen unidades oracionales propias de un área específica (órdenes informáticas, interjecciones deportivas de valor oracional, etc.).

situación comunicativa. Cabré (1999: 33), al explicar el carácter interdisciplinario de la terminología, afirma que la Terminología recibe los aportes de

> "... una *teoría del lenguaje* que dé cuenta de las unidades de significación especializada dentro del lenguaje natural teniendo en cuenta que participan de todas sus características, pero singularizando su carácter especializado y explicando cómo se activa este carácter en la comunicación..."

3. Existen regularidades en el comportamiento de las soluciones de traducción de este tipo de sintagma del inglés al español.

El estudio piloto del corpus paralelo nos permitió observar que a pesar de la variabilidad sintáctica de los patrones, existen regularidades que se deben verificar en un corpus de mayor tamaño. La sistematización de estas regularidades nos permitirá proponer soluciones que permitan a los traductores y los terminólogos solucionar de manera rápida y confiable este tipo de sintagmas. Tangencialmente, la sistematización de estos sintagmas nominales nos permitirá solucionar la mayoría de casos en traducción automática que, como Woolie (1997) establece, son los casos más complicados de solucionar y que más errores generan.

1.5 Objetivos de la tesis

Con base en las hipótesis de esta tesis, se pretende resolver básicamente dos problemas: uno teórico y uno aplicado.

En el plano teórico se pretende:

1. Demostrar que la existencia de los sintagmas nominales extensos especializados es una característica de la lengua que se presenta con mayor frecuencia en el discurso especializado.

En el plano aplicado se pretende:

2. Proponer recomendaciones para el tratamiento de estos sintagmas del inglés y sus correspondientes en español desde el punto vista formal y semántico para que profesionales de la traducción, la terminología, la lexicografía, la ingeniería lingüística, entre otros, puedan emplearlos en sus diferentes tareas profesionales.

Para cumplir estos dos objetivos generales, se propone una serie de objetivos específicos.

1. Analizar cuantitativamente los sintagmas nominales extensos especializados en ambas lenguas en un corpus escrito del ámbito de genoma.

2. Caracterizar formal y semánticamente los sintagmas nominales extensos especializados.

3. Observar si este fenómeno lingüístico es más productivo en el discurso especializado debido a la relación emisor-destinatario en un tipo de comunicación específico y natural de la lengua y en la relación de estas estructuras en la producción de conocimiento y la interpretación de los diferentes destinatarios involucrados (expertos y diferentes profesionales de las lenguas).

4. Diseñar un método de análisis de este tipo de sintagmas que permita interpretar estos sintagmas usando elementos lingüísticos resultado de las regularidades observadas en esta tesis.

5. Observar si existen diferencias en el uso de estos sintagmas con respecto a lo que produce un experto en cada lengua y lo que reflejan los textos paralelos (traducidos) y observar si son convenciones retóricas inherentes al inglés y al español.

6. Comparar el comportamiento de los sintagmas nominales extensos especializados en los diccionarios especializados con el de los corpus especializados.

7. Comparar el comportamiento de los sintagmas nominales extensos especializados en los corpus monolingües generales y especializados en español.

1.6 Organización de la tesis

En esta tesis no se sigue la estructura típica de todas las tesis debido a la diversidad de corpus que se emplean y a que se observan los aspectos sintácticos y semánticos de un fenómeno en cada lengua y posteriormente se contrastan con un corpus lexicográfico y luego se observan las regularidades de traducción en un corpus paralelo que, a su vez, es contrastado con los corpus monolingües.

Pensamos que presentar la metodología y todos los corpus en un sólo bloque sería muy extenso y confuso. De igual modo, los resultados y la discusión podrían diluirse si se colocaban en un solo bloque. Por tanto, hemos preferido dividir los aspectos metodológicos inherentes a cada lengua tanto para los aspectos sintácticos como para los semánticos. De igual modo, la metodología, los resultados y la discusión del análisis contrastivo se han separado en un capítulo.

De todos modos, se ha redactado una sección sobre aspectos metodológicos generales en los cuales se describen los corpus, las herramientas

de procesamiento, los problemas con el tratamiento de datos y la selección de las muestras.

La estructura general de la tesis se compone de una introducción en la que se presenta el objeto de estudio, las hipótesis y los objetivos de la tesis.

En el capítulo 2, se presenta el estado de la cuestión de los sintagmas nominales extensos especializados desde varias disciplinas. Se examina la confusión teórico-metodológica y la falta de criterios en muchos estudios. También se discuten las diferentes miradas que pueden hacerse de los sintagmas nominales extensos especializados y se contrastan los prejuicios existentes que están condicionando las descripciones y los trabajos teórico-descriptivos de los investigadores que trabajan en los estudios de traducción, en la enseñanza de lenguas para propósitos específicos y en otras disciplinas.

En el capítulo 3, se expone la metodología general de la tesis. Se describen los diferentes tipos de corpus, su procesamiento y las diferentes herramientas empleadas. Se explican los problemas con el etiquetaje sintáctico y semántico, y se exponen las decisiones tomadas para solucionarlos.

En el capítulo 4, se describen y analizan cuantitativa y lingüísticamente los patrones sintácticos del inglés. Se discuten los criterios de selección del corpus de análisis y presentan los resultados en cuanto a su longitud, frecuencia, categoría léxica predominante, relaciones de dependencia, aspectos morfológicos, entre otros. Se contrastan estos resultados con los resultados que se han obtenido del análisis de los diccionarios, de modo que se pueda constatar que los sintagmas nominales extensos especializados son un fenómeno de la lengua que está presente en los textos y diccionarios y que su análisis puede sistematizarse. Al final del capítulo, se hace una recapitulación de los resultados más relevantes.

Al igual que en el capítulo 4, en el capítulo 5 se describen y analizan cuantitativa y lingüísticamente los patrones sintácticos del español. Se

examinan los criterios de selección del corpus de análisis y se presentan los resultados en cuanto a su longitud, frecuencia, categoría léxica predominante, relaciones de dependencia, aspectos morfológicos, entre otros. Se contrastan estos resultados con los resultados obtenidos del análisis de los diccionarios. Es importante recalcar que en este capítulo se contrastan el corpus de análisis y los resultados cuantitativos del corpus CREA de la RAE para observar si las estructuras empleadas por los expertos concuerdan o no con las del lenguaje general. Al final del capítulo, también se hace una síntesis de los resultados más relevantes.

En el capítulo 6, se realiza el análisis semántico en inglés con WordNet 2.1 y UMLS 2006AB con el fin de observar las clases semánticas que predominan en los núcleos y la premodificación de los sintagmas nominales extensos especializados. A partir de estas clases semánticas se obtienen los patrones semánticos en cada programa y se examina cómo se correlacionan con los patrones sintácticos y el área de conocimiento. De igual modo, contrastamos los resultados de las categorías de WordNet 2.1 con UMLS 2006AB.

En el capítulo 7, se hace el análisis semántico en español con EuroWordNet 1.6 con el fin de observar las clases semánticas que predominan en los núcleos y la premodificación de los sintagmas nominales extensos especializados. A partir de estas clases semánticas se obtienen los patrones semánticos y se examina cómo se correlacionan con los patrones sintácticos y el área de conocimiento.

En el capítulo 8, se realiza el análisis del corpus paralelo en cuanto a longitud, patrones más frecuentes en inglés y las soluciones de traducción más frecuentes en español, categoría predominante, entre otros. Se analizan los patrones más frecuentes y sus relaciones de dependencia y se comparan con los patrones para observar las tendencias de cada corpus y si se cumplen las regularidades de los patrones. Estos patrones en inglés y sus patrones equivalentes en español se comparan contra los patrones del diccionario Mosby de medicina para observar las tendencias en las soluciones hacia el español.

Posteriormente se comparan los patrones obtenidos en español con los patrones del corpus de análisis en español, el corpus lexicográfico y el CREA de la RAE para poder observar sus tendencias y evidenciar si puede haber interferencias en las soluciones dadas por los traductores.

Finalmente, en el capítulo 9 se presentan las conclusiones, las limitaciones del estudio y las recomendaciones para el trabajo futuro en este tema.

2. Los SNEE vistos desde diferentes disciplinas del lenguaje

2.1 Introducción

Los sintagmas nominales extensos especializados (SNEE)[4] se han visto desde diferentes disciplinas relacionadas con el lenguaje: gramática clásica, lingüística teórica, lenguajes especializados (LSP), terminología, traducción, traducción automática, inglés para propósitos específicos (ESP), escritura técnica, recuperación de la información, entre otros. Sin embargo, no hay una descripción detallada de ellos hasta donde se ha explorado.

Muchas disciplinas ven estos sintagmas más como un obstáculo o una rareza de la lengua, como se verá a continuación, más que como un fenómeno recurrente y "natural" en ciertos niveles de especialidad o géneros discursivos como se ha ilustrado en §1.3.

Es importante tener en cuenta que los fenómenos lingüísticos no deben considerarse únicamente bajo una óptica prescriptivista como se verá más adelante, pues esta visión impide explicar por qué ocurre un fenómeno de este tipo. Hay que tener en cuenta que son los científicos los que usan este tipo de sintagmas nominales y no los lingüistas y que debe existir otro tipo de motivaciones para usar determinadas características de la lengua con más frecuencia y de algún modo específico.

2.2 Un fenómeno, muchos nombres

Se pueden distinguir básicamente dos tipos de denominación para los sintagmas nominales extensos dependiendo de sí la unidad está lexicalizada o

[4] En adelante nos referiremos también como sintagmas nominales (SN).

no. En muchas disciplinas las denominaciones se acuñan independientemente del grado de lexicalización del sintagma. Sólo los terminólogos hacen una distinción clara, por ejemplo, entre unidades términos o unidades libres (Cabré 1993), y actualmente, entre términos, fraseología, colocaciones, locuciones y unidades libres (Lorente 2001).

Kocourek (1979) utiliza la denominación *lexical phrase* y para los nombres *compounds* (como parte de un *lexical phrase*). Igualmente, emplea *French noun-phrase terms*. Como puede verse, los nombres corresponden a unidades que pueden entrar en el lexicón de una lengua y no a unidades que pueden ser fruto de un encadenamiento accidental dentro del discurso.

lexie	(Rey 1977: 15; Pottier 1973: 251)
synapsie	(Benveniste 1966: 91)
paralexème	(Dubois *et al* 1973: 354; Griemas 1966: 37-38)
mot complexe	(Auger 1975: 79, Picoche 1977: 23)
syntagme codé	(Rey 1977: 135)
groupe lexical	(Vachet 1964: 40)
lexie complexe	(Pottier 1963, 1968: 19, 1973: 251)
lexème complexe	(Goffin 1978)
groupe lexicalisé	(Rey 1975: 13)
locution composée	(Bally 1975: 250)
syntagme lexical	(Auger 1978)
syntagme autonome	(Guilbert 1967: 305)
syntagme lexicalisé	(Rey-Debove 1971: 113; Rey 1975: 11)
lexème syntagmatique	(Lyons 1978: 25-26)
composé syntagmatique	(Auger 1976: 66)
dénomination complexe	(Hollyman 1966: 97)
dénomination synaptique	(Benveniste 1966: 94)
groupe de mots	(Phal 1964: 47)
syntagme de lexique	(Auger 1976a: 67)
unité de signification	(Guilbert, Quemada 1972: 400)
unité lexicale complexe	(Phal 1969: 76)
unité sémantique complexe	(Dubois 1960: 62-63)

unité lexicale supérieure	(Marcellesi y Marcellesi 1969: 114)
unité syntagmatique de signification	(Guilbert 1967: 305)
unité lexicale syntagmatique complexe	(Dugas 1978).

Kocourek excluye la siguiente denominación:

unité lexicale à deux et plus de deux éléments

Esta primera clasificación de unidades terminológicas extensas para el francés la realizó Kocourek (1979: 124) que luego la publica en 1981 y 1991. Kocourek (1979) no tiene en cuenta para este listado los nombres en francés que también tienen sintagmas no lexicalizados.

Sin embargo, otros autores ponen en duda el carácter lexicalizado de algunas unidades y por esto, prefieren tomar un rango más amplio de estas unidades. Además, dependiendo del área de estudio como en la traducción automática, los estudios de traducción y la enseñanza del inglés con propósitos específicos, el hecho de que una unidad esté o no lexicalizada pasa a un segundo plano como lo plantea L'Homme (1994: 150):

> "La distinction souvent évoquée, notamment en terminologie, entre un groupe lexicalisé et une combinaison libre joue un rôle secondaire de point du vue de la TA (traduction automatique). Les notions de lexicalisation ou de véhicule d'une signification ou d'une notion unique ne sont pas vraiment mises à contribution".

Esto se debe a que en estas disciplinas un sintagma nominal de estas características es un problema que se debe resolver en el acto: analizarlo y traducirlo o interpretarlo.

En los manuales de gramática y en los artículos de corte más lingüístico no se denominan los sintagmas nominales extensos especializados con un nombre específico como se suele hacer en otras áreas; siempre son sintagmas

nominales con premodificación compleja o un tipo de compuesto de dos o más sustantivos o adjetivos:

noun phrases with multiple premodifiers	(Quirk *et al* 1985)
compound nouns	(Downing 1977: 810)
complex nominal	(Levi 1978)
noun strings, piled-up adjectives	(Georges 1996)
multiple premodification	(Biber et al 1999)
sinapsia	(Alvar 1993: 22).

Esto se debe quizá a que se quiere potenciar más las partes que el todo. Por ejemplo, se dedican fragmentos de estas gramáticas a explicar las relaciones internas de los compuestos de dos elementos (*compounds*), los tipos de premodificación, el orden de los premodificadores, los tipos de núcleos, etc.

Por el contrario, en terminología, como es de esperarse, predominan las denominaciones relacionadas con las palabras término y terminología, es decir, sintagma terminológico y término sintagmático; se considera que siempre están lexicalizadas. Para identificar la cantidad de palabras involucradas en el sintagma se usan expresiones como multi-, poli-, y complejo. Sin embargo, la palabra 'complejo' no siempre indica cantidad de palabras en algunos autores.

complex noun phrases	(Myking 1989)
secuencias de dos o más palabras	(Artnz y Picht 1989: 150)
grupos léxicos rotgráficos	(Wüster)
syntagmes termes	(Portelance 1989: 400)
syntagmes nominaux complexes (lexical o no)	(Jastrab 1987)
unités terminologique complexes	(Assal y Delavigne 1993)
sintagmas terminológicos	(Cabré 1993: 29)
terme complexe	(Bourigault 1993)
multiword compounds	(Maalej 1994: 142)
syntagme terminologique	(Collet 1997: 2003)
términos sintagmáticos	(Cartagena 1998: 282)
sintagmas terminológicos	(Aldestein 1998)

sintagmas terminológicos	(Cardero 2000)
unités terminologiques complexes	(Café 1999)
unités terminologiques polylexicales	(Estopà 2001: 219)
unidades terminologica complexa	(Faultisch 2003)
término sintagmático, término compuesto, complex term,	
grupo nominal complejo	(Méndez 2002: 191)
sintagma nominal compuesto	(Oster 2003: 138)
syntagme nominal pertinents	(Naulleau 1998: 33).

En traducción, no se usan, en muchos casos, nombres para designar el fenómeno como en el caso de López y Minett (1997), pero se intenta dar indicaciones para resolverlo o comentarlo. Como consecuencia, no se ve que predomine un nombre en la literatura de traducción aunque puede afirmarse que autores como Newmark para el inglés y Alcaraz para el español se pueden tomar como puntos de referencia.

multiple noun compounds	(Newmark 1981: 115, 1988: 41)
sintagma nominal largo	(Alcaraz 2000)
noun clusters	(Linder 2002)
multinoun compounds	(Quiroz *et al* 1997, 2000)
sintagma nominal extenso especializado	(Quiroz *et al* 2004)
sintagmas nominales complejos	(Zabala 1996)
compuestos multinominales	(Abril y Ortiz 1998: 291)
grupo nominal complejo	(Montero 1995: 50)
yuxtaposición de adjetivos	(López y Minett 1999: 105)

En la escritura técnica es más usual hablar de cadenas de sustantivos (o de adjetivos denominales). A partir del uso de algún adjetivo de manera peyorativa, para criticar el exceso de premodificación, tal como se evidencia en los autores siguientes:

nounspeak	(Orwellian 1974)

noun strings	(Burnett 1992: 312; Huckin 1991)
sustantivos adjetivales en “caravana”	(Norman 1999)
string/excessive premodification, groups of premodifiers	(Kirkman 1992)
nouns as adjectives	(Blake y Bly 1993)
complex nominals	(Montero 1996: 58)

En traducción automática e ingeniería lingüística es más importante la extensión del sintagma por sus posibilidades de interpretación que el hecho de que la unidad esté lexicalizada. Además, no sería rentable para un sistema de TA incorporar en su lexicón estas unidades tan extensas. En general se piensa que es preferible tener un analizador sintáctico (*parser*) que permita hacer un análisis de los sintagmas nominales ya que éstos son recurrentes en los textos técnicos.

compound structures	(Chambers 1994)
noun sequences	(Lehrberger 1982: 29)
nominal compound	(Finin 1980: 1986)
compound nouns	(Bennett 1993: 43)
groupes nominaux	(L’Homme 1994: 148)
noun compounds	(Barker 1998)
compound nominal groups	(Woolley 1997)
two-noun compounds/nominal compounds	(Maalej 1994)
noun sequences, noun compounds, complex nominals	(Vanderwende 1995: 2)
noun + noun compound	(Downing 1977)
noun sequence	(Leonard 1984)

En ESP tiende a predominar la palabra compuesto, pero no en el sentido estrictamente binario y lexicalizado que ven los lingüistas. Por esto deben agregarse palabras como ‘*complex*’, ‘*phrase*’, ‘*in chain*’, ‘*group*’, *etc.* para dar una idea de una unidad más larga de lo “normal”.

noun strings	(Palmer 1968)

noun compounds (terms)	(Pugh 1984: 395)
long noun phrase	(Varantola 1985)
complex compounds,	
more complex compounds,	
very complex compounds,	
noun strings	(Trimble 1985)
compound nominal phrases	(Salager 1984)
nominal compounds	(Horsella y Pérez 1991: 125)
lessie complesse	(Casadei 1994: 58)
nominal groups	(Thouvenin 1996)
complex nominals	(Montero 1996)
nominal group	(Ventola 1996)
compound nominal groups	(Woolley 1997)
groupes nominaux	(Maniez 2001)
noms composés en chaîne,	
enchaînements syntagmatiques complexes	(Boughedaoui 1995; 2001: 139)
syntagmes nominaux complexes	
syntagme nominal fleuve	(Le Masle 2001: 65)
noms composés (lexicalisés),	
groupe nominal complexe	(Ormond 2001: 9)
noun clusters	(Limaye Pompian 1991; Salager 1985)

Puede verse en este breve recorrido por diversos autores que este fenómeno presenta diversas interpretaciones, dependiendo del área y por esto la variedad de denominaciones del fenómeno. Esto hace su abordaje más complejo, pero a la vez más enriquecedor.

2.3 Gramática clásica

Las gramáticas clásicas del inglés y los lingüistas poco se refieren a los sintagmas nominales extensos, sólo algunos autores como Quirk *et al* (1985) y Biber *et al* (1999) mencionan el fenómeno. Estos últimos autores presentan un

análisis estadístico por niveles de lengua (registros) y explican los casos internos de coocurrencias en la premodificación de algunos patrones. Otros autores como Levi (1978) mencionan este tipo de sintagma nominal, pero no hacen un análisis de las estructuras extensas. Esta autora deja implícito que los sintagmas nominales con adjetivos no predicativos (denominales) de dos unidades (tipo compuesto) pueden servir para analizar los más extensos.

Los gramáticos clásicos (Quirk *et al* 1985; Huddleston y Pullum 2002) ven estas unidades de diferentes modos.

De entrada, Quirk *et al* (1985: 1338) ponen de manifiesto que hay suficiente evidencia para decir que el uso de esta característica es más frecuente en la forma escrita que en la oral.

> "There is indeed evidence of a higher proportion of three-or-more item sequences in written than in spoken English"

No obstante, Quirk *et al* (1985: 1342), cuando explican la premodificación múltiple dan cuenta de los SN como "cosas raras" e improbables.

> "It should be noted, however, that if we introduce an adjective in this last noun phrase, already *clumsy* and improbable, the adjective has to come…" (itálicas nuestras).

Además, estos autores plantean que los sintagmas nominales con premodificación compleja son un medio eficiente de comprimir la información, pero que implicitan las relaciones semánticas entre los modificadores, lo que puede ocasionar problemas de interpretación para los hablantes.

Quirk *et al* (1985: 1342-43) establecen que la premodificación múltiple presenta básicamente dos problemas: 1) orden relativo y 2) límite psicológico.

Cuando hay más de un premodificador surgen problemas de orden. Analizan que el orden que hay en la premodificación lineal de:

[Expensive [overseas [income [tax [office furniture]]]]][5]

no es común y, por el contrario, pueden surgir varias interpretaciones como en:

A [new [giant size] [cardboard [detergent carton]]]

donde 'size' no premodifica a 'cardboard' y 'cardboard' no premodifica a 'detergent'.

Quirk *et al* (1985) y Biber *et al* (1999) reconocen que este fenómeno no es algo oscuro, pobre o excesivamente extenso. Por el contrario, piensan que esto depende del tipo de destinatarios a quien va dirigido el texto. Piensan más en términos de implicitación de las relaciones semánticas para quienes no son expertos:

> "such an example is not, of course, obscure. Indeed, it is generally the cause that obscurity in premodification exists only for the hearer or reader who is unfamiliar with the subject concerned who is not therefore equipped to tolerate the radical reduction in explicitness that premodification entails" (Quirk *et al* 1985: 1343).

Otros teóricos de la lingüística, en especial las gramáticas clásicas del inglés han detectado el fenómeno a partir de la descripción de los patrones de concurrencia más frecuentes (hasta 4 tokens) para la lengua inglesa y en varios géneros textuales. Sin embargo, no han estudiado a fondo este fenómeno en el

[5] Aunque pensamos que debería interpretarse [expensive [overseas [income tax]] office furniture]].

tipo de texto que se emplea en este estudio. (Quirk *et al* 1985: 1338; Biber *et al* 1999: 595-597; Huddleston y Pullum 2002: 453).

De estos autores, Biber *et al* (1999) han explicado de manera general y cuantitativa la premodificación compleja de hasta 4 tokens de 4 tipos de patrones y las relaciones semánticas de tipo compuesto (como '*compounds*') dentro de la premodificación en varios géneros discursivos: lengua general, noticias y prosa académica, en especial.

Para estos mismos autores (1999: 589), la premodificación múltiple es más común en las noticias que en la prosa académica, pero ambas en conjunto tienen una frecuencia más alta que en el resto de registros[6]. Además, proporcionan unos patrones de coocurrencia de la premodificación de los sintagmas nominales y muestran que la premodificación es más común que la posmodificación en el registro académico:

> "In all registers, NP (noun phrases) with premodifiers are somewhat more common than those with postmodifiers (578). Proportionally, in academic prose, almost 60% of all NP have some modifier of which 25% have a premodifier and 20% have a postmodifier" (Biber *et al* 1999: 579).

Como bien lo dicen Biber *et al* (1999: 579) se explica que hay más sintagmas nominales con premodificación que con posmodificación debido a que la nueva información se empaqueta a modo de modificadores en el sintagma nominal:

> "Much of the new information presented in academic texts is packaged as modifiers in NP (noun phrases), resulting in a very high density of information".

---

[6] En este trabajo se trata de usar la terminología empleada por cada autor. Registro se usa en el sentido de niveles de lengua. Concuerda, en cierto sentido, con nivel de especialidad.

Sin embargo, Abberton (1977: 29-72) establece lo contrario con base en un corpus de ocho textos del corpus de *Survey of English Usage*, cuatro de novelas y cuatro de escritura científica no popular (*non-popular scientific writing*). De estos últimos, un texto pertenece al área de la biología, dos a la física y la matemática, y uno a la química. Abberton concluye que:

> "both types of English examined are remarkably similar (subrayado del propio autor) in nominal group premodification structure: the vast majority of nominal groups are not premodified or are premodified by only one determiner and/or adjective (…) the main differences between the fiction and the science texts are that in the scientific writing postmodified nominal groups predominate; …The spread over form class types is similar in both styles of writing and there is no form class (pattern) exclusively preferred by one or the other. In both the fiction and the science texts the 'favourite' patterns for premodified nominal groups are the same, although postmodified nominal groups are far more common in the science texts" (Abberton 1977: 62-63).

El hecho de que predomine la posmodificación sobre la premodificación según esta autora, se debe principalmente al tipo de corpus (registro de la lengua) y quizás al área del conocimiento en cuestión. Sin embargo, casi todos los corpus del inglés tienen nuestras áreas en cuestión: biología, química, física, etc. Una revisión rápida de estos corpus nos muestra que no hay o hay muy pocos textos del registro o nivel realmente (muy) especializado. Es decir, que hay pocos textos (y cantidad de palabras) que tengan una situación comunicativa del nivel experto-experto. Con lo cual, si se tiene en cuenta la cantidad de textos que se producen en el ámbito científico-técnico, las generalizaciones que se hacen en algunos casos no son representativas ni describen situaciones comunicativas reales. Además, como se ve en la cita, este estudio contradice los datos en cuanto a la cantidad de pre- y posmodificación en los textos especializados. Puede verse que aún en estudios con corpus hay generalizaciones contradictorias. Quizá en el caso de este estudio (1977), se debe al tipo y a la cantidad de corpus empleado. Además, no existían muchos avances

que permitieran tener grandes cantidades de corpus ni las herramientas que hoy se tienen.

Por otro lado, Biber *et al* (1999: 589) establecen que los premodificadores son más comunes en las noticias que en la prosa académica. Sostienen que también los adjetivos comunes (i. e., adjetivos no participios) son la categoría gramatical que, como premodificador, es más común en todos los registros:

> Los adjetivos premodificadores son extremadamente comunes en la prosa académica.
>
> Los sustantivos representan el 40% de los premodificadores en las noticias y un 30% en la prosa académica.
>
> Los modificadores en *–ed* son algo más comunes en la prosa académica que en otros registros.

Asimismo, estos autores establecen que "only about 2% of premodified noun phrases have 3 or 4 word premodification", sumando todos los registros del corpus Logman. Además, argumentan que el género de noticias tiene un poco más de premodificación compleja: "In news longer premodification sequences are slightly more common". Sin embargo, no se explican las relaciones internas de la premodificación con su núcleo en los sintagmas nominales de más de tres tokens.

Puede verse que no hay un acuerdo en los datos pues casi todos dependen del tipo de corpus como se dijo antes. A pesar de esto, se puede observar que este tipo de unidades tan extensas no tiene una frecuencia alta en la lengua en su totalidad, pero si en otros registros o géneros de la lengua inglesa.

Al hablar de la premodificación Quirk *et al* (1995: 1337-1338) dicen que:

> "the problem becomes even more acute with longer strings of premodifiers. Although there is, theoretically, no grammatical upper limit to the number of

premodifiers, *it is unusual to find more than three or four*" (las ítálicas son nuestras).

No obstante, estos mismos autores matizan que "premodification is an area of English grammar where there is considerable variation among the varieties of the language". Esto quizá se debe al tipo de corpus que usan y la baja frecuencia que ellas tienen en corpus heterogéneos como los usados en dichas gramáticas (Quirk *et al* 1995: 1337-1338).

Algunos autores como Levi (1978) y Woollie (1997) establecen que los análisis bajo la Teoría de la Barra X presentan problemas para tratar los atributos y complementos de un sintagma nominal, quizá por la cantidad de irregularidades que pueden presentar.

Dada la capacidad de la Gramática Sistémico-Funcional de Halliday para tratar la premodificación de un sintagma nominal como *deitic + numerative + epithet + classifier + head/thing*, algunos autores han analizado la premodificación de los sintagmas nominales bajo esta misma perspectiva (Thouvenin 1996; Boughedaoui 2001: 138; Ormod, 2001: 12). De este modo, se puede distinguir entre los adjetivos que funcionan como atributos del N y los que funcionan como complemento de la posmodificación. Igualmente, Woollie (1997) afirma que se ha estudiado poco la estructura interna de los sintagmas nominales extensos.

2.4 Terminología

Desde el punto de vista terminológico y de los lenguajes de especialidad se ha descrito muy poco la estructura interna y el comportamiento de los sintagmas nominales extensos en los textos especialidad, no desde el punto de vista de la lexicalización sino como unidades (libres) de los discursos especializados que condensan una gran cantidad de información en poco

espacio. Así el uso de este tipo de estructura puede, en un momento dado, establecer la densidad conceptual de un texto y permitir, por tanto, clasificar un texto por su nivel de especialización.

Estas unidades involucran una serie de problemas como el comportamiento de los elementos verbales y no verbales, v. g., siglas, acrónimos, formas cortas, inclusión de siglas en otras siglas, abreviación discursiva (en una revista sí y en otra no), cifras, códigos, entre otros fenómenos que deberían ser explicados. Otro aspecto importante sería observar el grado de lexicalización de dichas unidades y saber si tienden a ser unidades terminológicas especializadas o son simplemente un conjunto de ellas o una mezcla de lenguaje general y lenguaje especializado (discurso especializado) mediante la repetición de estas unidades en otros textos.

Tal y como se ilustró en los ejemplos de los tres diccionarios descritos en §1.3, los sintagmas nominales especializados extensos están presentes en los diccionarios técnicos.

Al hablar de la lexicalización de los sintagmas terminológicos Kocourek (1991: 140) plantea que dichas unidades tan extensas no pueden ser nombres sino definiciones y, además, carecen de estatus léxico: "… une expression telle que *liqueur alcalique saturée de la matière colorante du bleu de Prusse* est 'moins que un nom qu'une définition'…" En esa misma línea, Cartagena (1998: 281-296) concuerda con Kocourek en cuanto a que "(desde luego que) existe una relación directa entre longitud, el grado de especialización y la estabilidad sintáctica del término; a mayor longitud, mayor especialización e inestabilidad." Arntz y Picht (1989: 150) también ponen en duda el carácter terminológico de dichas estructuras. En este sentido, Cabré (1993: 304) dice que para los sintagmas muy extensos se suelen utilizar formas reducidas:

> "Comunicativamente, los sintagmas terminológicos excesivamente largos se suelen utilizar de forma abreviada en el discurso, y, a la larga, se suelen resolver fijando las formas reducidas respectivas como términos de uso corriente:

unidad central de proceso:	unidad central
hoja de cálculo electrónico:	hoja de cálculo
red de área local:	red local”

Cartagena (1998) agrega también que no se ha estudiado los términos sintagmáticos en el ámbito del LSP para el español. De igual modo, establece unos patrones más regulares en español de hasta 4 tokens en el ámbito de la anatomía. A continuación, se presentan los patrones de tres y cuatro modificadores para el español obtenidos por Cartagena (1998: 283-284).

N Adj Adj Adj (el más frecuente)
Adj N Adj Adj (el más frecuente)
N Prep D N Adj Adj
N Prep D N Adj Prep D N
N Prep N Prep N Prep N
N Adj Prep D N Adj
N Adj Prep D N Prep N
N Adj Adj Prep D N

N Adj Adj Adj Prep D N
N Adj Adj Prep D Adj N
N Adj Prep D N Adj Adj

En este estudio, Cartagena presenta los patrones para explicar la variabilidad de los sintagmas terminológicos. Él no presenta una estadística de su frecuencia respecto del corpus, simplemente se limita a decir las veces que aparece dicha estructura.

Cartagena (1998) concuerda además con autores como Cabré (1993) y Hoffmann (1985) en cuanto a que la combinación de los componentes de los sintagmas terminológicos está regida por la gramática de la lengua común.

> "La sintagmación, como recurso formal de obtención de unidades léxicas, se basa en la formación de una nueva unidad a partir de una combinación sintáctica jerarquizada de palabras. Las nuevas unidades así obtenidas respetan las reglas combinatorias del sistema lingüístico al que pertenecen, e incluyen muy frecuentemente conectores gramaticales…" (Cabré 1993: 177).

Como se expresó antes, en este trabajo se mantendrá la postura de que no existe una dicotomía entre lenguaje general y "lenguaje" especializado (o mejor dicho discurso especializado); no es un problema de "otro" lenguaje con características excluyentes. Se afirma mejor que la descripción y el análisis de los patrones se harán desde la óptica de una gramática general y, su uso y frecuencia en el discurso especializado se deben a razones conceptuales o cognitivas.

En español, hay pocos estudios desde el punto de vista terminológico que incluyan unidades tan grandes. Cardero (2000, 2004) hace una descripción de los sintagmas terminológicos en un corpus de cinco áreas: cinematografía, el Tratado de Libre Comercio, redes de computación, telefonía celular y control de satélites. Esta autora no proporciona el número total de palabras del corpus ni el método de extracción. Además, no establece el tipo de corpus: textual o terminológico; parece ser que es mixto. Del total de términos, selecciona aleatoriamente 1.368 unidades (20% del total). De estos, el 5,19% corresponde a 66 términos de 12 estructuras que la autora denomina poco frecuentes. Estas estructuras se dan principalmente en el área de la telefonía celular. Si bien es un estudio pionero en este tipo de unidades, no profundiza sobre aspectos formales ni semánticos.

Dentro del marco de la TCT, Estopà (1999, 2001) hace una clasificación de las unidades terminológicas especializadas de acuerdo con las preferencias de los usuarios. Establece que las unidades que los traductores seleccionan son segmentos de las unidades terminológicas poliléxicas (principalmente nominales y adjetivales). Según ella, esto se debe a que son las que ocasionan problemas durante el proceso traductivo.

A pesar de la clasificación de las unidades, la autora no analiza unidades tan grandes; las estructuras de más ocurrencias tienen un núcleo con dos modificadores (dos complementos) del tipo N Adj Adj = 30 ocurrencias. De su tesis se pueden comparar con nuestro estudio los datos de tipo morfológico y observar si las estructuras que la autora plantea (1999: 114) se presentan en nuestro corpus y, si estas simplemente se expanden o se forman otras estructuras no descritas. En un trabajo previo (Quiroz 2004) se han comparado sintagmas desde dos premodificadores y, se concluye que una de las estructuras más frecuentes en inglés es Adj Adj N y todas las soluciones en español están representadas en la estructura N Adj Adj.

Sin embargo, en el mismo trabajo otras estructuras recurrentes como Adj N N = N N Adj, Adv PP Adj Adj N = N Adj Adj Adj Adv, entre otras, no están descritas en Estopà (1999). Debe aclararse que en Quiroz (2004) se han extraído sintagmas nominales lexicalizados y no lexicalizados. Por tanto, se debe mirar con cuidado los datos para no hacer generalizaciones inadecuadas. A pesar de esto, todas las estructuras lexicalizadas ya descritas en otros trabajos (Cartagena 1998; Estopà 1999; Café 2000; Cardero 2000; Vivaldi 2004) se tendrán en cuenta puesto que para la traducción son igualmente relevantes. Es importante destacar que Vivaldi (2004: 2-3) recoge los patrones más frecuentes de sintagmas nominales especializados, que denomina patrones terminológicos complejos: N Prep N Adj, N Adj Prep N Adj, N Adj Adj y N Adj Adj Prep N Prep N.

Dentro del marco de la fraseología contrastiva en el ámbito de la radiología, Méndez (2000, 2002) incluye los grupos nominales complejos dentro de las estrategias discursivas de este tipo de discurso. Méndez (2002) se centra en una serie de estructuras muy diversas a partir de las concordancias extraídas con el programa WordSmith. Por tanto, no hay análisis formal (patrones), semántico ni estadístico de esta variedad de estructuras que incluya, entre otros, los sintagmas nominales. El objetivo de Méndez es la orientación pragmática de la fraseología para traductores y redactores. Sin embargo,

además de las estrategias fraseológicas que pueden emplear los traductores o redactores, también necesitan las regularidades en el plano lingüístico para tener estrategias claras para resolver los sintagmas nominales con independencia de los elementos que los acompañen en la predicación. Así, pues, las relaciones internas del sintagma son más relevantes para esta tesis, sin dejar de reconocer la importancia de las otras estrategias, que tienen que ver más con elementos de coherencia y cohesión, como esta autora lo plantea.

Finalmente, Café (2000) presenta una clasificación de la expansión y las reglas de formación de las unidades terminológicas complejas para el portugués de Brasil en el ámbito de la biotecnología desde la perspectiva de la gramática funcional de Dik. Café parte también del principio de que una gramática de la lengua general debe abarcar los llamados lenguajes de especialidad y, por tanto, para crear las reglas de expansión se basa en una teoría de la predicación. Las reglas están compuestas por una base (el núcleo), un argumento (el complemento) y los satélites (complementos externos). Cada base, argumento o satélite tiene una función semántica y una función sintáctica (sujeto, complemento nominal, complemento circunstancial, etc.). Para los propósitos de esta tesis, algunos elementos del análisis de Café son interesantes, v. g., las funciones semánticas de un ámbito similar al nuestro. Sin embargo, puede considerarse que esta autora hace un análisis forzado al adaptar unas funciones de la predicación al ámbito de los sintagmas nominales, cuestión que pertenece al ámbito del verbo. Además, este estudio no contiene un análisis estadístico, un análisis de relaciones semánticas ni un análisis contrastivo entre dos lenguas como se pretende en esta tesis. El análisis de los satélites de los sintagmas nominales que hace Café no tiene una función clara dentro de las reglas de formación pues no los incluye. Por un lado, la regla de formación más extensa da cuenta de sintagmas nominales con una extensión máxima de tres modificadores del núcleo (*molécula de DNA circular extracromossómica*).

2.5 Estudios de traducción

Desde la traducción, algunos autores como Vázquez-Ayora (1977), López y Minett (1997), Zabala (1998), Linder (2001 y 2002), entre otros, afirman que los sintagmas nominales extensos son un problema de traducción dado el problema que plantea el orden de los premodificadores y la falta de relaciones explícitas de las relaciones semánticas entre los diferentes elementos. Además, su traducción en las lenguas romances como el español, el catalán o el francés presenta muchas posibilidades para explicitar las relaciones a través de la posmodificación.

Ya en 1959, Vinay y Dalbernet dedican un apartado a explicar los problemas que los sintagmas nominales y los compuestos ocasionan al traductor en la traducción general y advierte de los peligros que ellos albergan (1959: 152-153).

De igual modo, Newmark (1988: 40) establece *a priori* que los compuestos con múltiples sustantivos ("multinoun compounds") son una característica de los textos formales que se suelen traducir (según su tipología). Quiroz *et al* (1997) muestra en un corpus pequeño en el área de las enfermedades tropicales que los sintagmas nominales extensos son una característica relevante de los textos especializados.

Abril y Ortiz (1998: 291) centran el problema de los sintagmas nominales extensos dentro de las características generales del lenguaje médico y, en especial, desde el punto de vista gramatical:

> "Así, en inglés médico son habituales la nominalización, los verbos en pasado, el participio pasado, el participio presente, los compuestos multinominales (*mitogen-triggered lymphocyte DNA synthesis*)..."

Sin embargo, estos autores no proponen un análisis sintáctico-semántico o de otra índole para solucionar el problema. Las propuestas, cuando las hay, son muy intuitivas y no responden a una lógica sistemática de estos casos. Algunos de ellos como Linder (2002) también los ven como un problema.

Uno de los pocos autores que tiene una perspectiva similar a la nuestra es Cartagena (1998: 282) dentro del marco de variabilidad de términos para el par alemán-español en el ámbito de la anatomía. Cartagena comenta que algunos autores como Coseriu (1973: 11) afirman que este tipo de unidades no presenta dificultad alguna para la traducción ya que no se trata de transposición de significados sino de un mero reemplazo de significantes en relación uno a uno. Calonge (1995: 184-185) también argumenta que parece evidente que el vocabulario científico no tiene nada que ver con la lengua general [...] los sintagmas que representan significados especializados [...] en general son fácilmente traducibles, debiendo evitarse solamente el calco. Cartagena se apoya en su estudio y, en Artnz (1982: 114-117), para rebatir estos argumentos y mostrar que los sintagmas terminológicos presentan considerables problemas de traducción, incluso aquellos sintagmas nominales relativamente sencillos. Además, Cartagena muestra que los sintagmas terminológicos y su variabilidad responden a la sintaxis de la lengua general.

Desde una perspectiva prescriptivista, Vázquez-Ayora (1977: 123) establece que la manera lógica de resolverlos es: “to translate the premodifier closest to the nucleus and continue from there translating each successive adjective to the left and so on”. Este tipo de solución puede dar resultados en algunos casos, pero no es la única solución; sólo cubriría un tipo de patrón que quizá no sea el más frecuente.

Sin establecer unas regularidades o una lógica, López y Minett (1997: 103-109) proponen que deben analizarse los elementos adjetivales y reordenarlos en grupos alrededor del núcleo del sintagma. Linder (2002: 266) establece una serie de parámetros y dice que se debe, en primer lugar, buscar el núcleo, en segundo lugar, determinar el orden e importancia relativa de los

elementos y finalmente, aplicar la siguiente estrategia (sin un orden predeterminado): maximizar el número de elementos en español, variar las preposiciones, omitir sustantivos o partes del sintagma nominal si se repiten dentro de él (variación denominativa) y usar tantas técnicas de transposición como se pueda (p. ej., de sustantivo a verbo).

Vivanco (1994: 755) también sugiere una estrategia similar a la de los anteriores autores: "la traducción al español de estos grupos nominales, comienza por el sustantivo del final, ya que es a éste al que califican todas las demás palabras" y concluye que

> "Como norma general, se traducen los demás nombres y las formas $-ing$ y $-ed$, encontrando los equivalentes precisos en castellano e introduciendo las preposiciones que sean necesarias en español".

Puede verse que la estrategia tiende a ser prescriptivista y no responde a una lógica de análisis lingüístico o traductivo. En primer lugar, no explica cómo se debe determinar el orden y la importancia relativa de los elementos del sintagma nominal al no analizar las tendencias en un corpus textual o terminológico. En segundo lugar, proponer una estrategia para omitir sustantivos del sintagma nominal es muy peligroso puesto que un sintagma nominal se puede diferenciar de otro sólo por un sustantivo que haga referencia a otro concepto o a un concepto que funcione como hipónimo o hiperónimo en la jerarquía. Además, en la traducción especializada esto podría considerarse como una falta de coherencia en el uso de un término, lo que podría ocasionar problemas de cohesión. Finalmente, no se puede maximizar el número de elementos sin razón alguna. Además, el uso de una preposición en una lengua no responde a una elección caprichosa del hablante sino que responde a una serie de restricciones gramaticales, pragmáticas, y en especial semánticas.

La extracción y la paralelización de sintagmas nominales extensos llevada a cabo en Quiroz *et al* (2004), permitió hacer una primera exploración sobre las regularidades que ellos albergan para refutar, a pequeña escala, las propuestas

de los autores antes mencionados. Como luego se verá, ninguno de estos autores ha trabajado observando regularidades en un corpus.

Finalmente, pensamos que existen regularidades en los textos que nos pueden dar luz para abordar mejor la traducción de estos sintagmas nominales, que se deben identificar las relaciones semánticas y que los corpus pueden ser útiles para observar las tendencias en las soluciones al español.

2.6 Traducción automática

En el ámbito de la traducción automática (HAMT) existen trabajos que han abordado el problema (Woolie 1997; Lehrberger 1982; L'Homme 1994; Maalej 1994). Tal es el caso de Woollie (1997) quien detectó que el principal problema de los traductores automáticos comerciales actuales es la traducción de sintagmas nominales complejos (de longitud variable), lexicalizados o no que no están incluidos en el lexicón de excepciones del sistema. Para solucionar este problema se tienen dos alternativas. La primera es tener los sintagmas nominales en el lexicón del sistema, lo cual es muy inviable, poco productivo y poco predecible en términos de la variabilidad de los textos y áreas temáticas. La segunda es estudiar con profundidad los patrones sintácticos y las relaciones semánticas entre los diferentes componentes del sintagma nominal para observar las regularidades y así introducir un conjunto de reglas que permita al sistema solucionar correctamente los sintagmas nominales.

El problema radica en que la formalización de estos sintagmas nominales es muy difícil por las supuestas irregularidades que se pueden encontrar. Por consiguiente, los errores causados por el orden de los premodificadores en un sintagma nominal complejo en las gramáticas formales hacen que los analizadores morfosintácticos de los programas de traducción automática del inglés hacia una lengua romance produzcan unos resultados desastrosos (Woollie 1997: 3).

De esta dificultad planteada para formalizar estos sintagmas nominales se derivan problemas para la traducción automática. Lehrberger (1982: 92-94) también estableció algo similar para el sistema de traducción automática TAUM que traduce informes meteorológicos del inglés al francés. Este mismo sistema estaba proyectado para traducir manuales de aviones en los cuales los sintagmas nominales especializados desempeñarían un papel importante. Ya que este "sublenguaje" (como lo llama el propio autor) es muy restrictivo, Lehrberger estableció 50 relaciones sintáctico-semánticas suficientes para las combinaciones posibles de los manuales[7]. Sin embargo, no se hicieron predicciones para la lengua general (1982: 94)

Igualmente, Montero (1996) hace un pequeño estudio usando un corpus terminológico de 4.235 términos de 2 a 5 tokens y concluye igualmente que la traducción de los sintagmas nominales extensos es una de las dificultades mayores y más visibles de los traductores automáticos.

2.7 Inglés para propósitos específicos (ESP)[8]

En la enseñanza del inglés con propósitos específicos (ESP), los sintagmas nominales extensos especializados son un verdadero problema para el profesor de lengua inglesa a la hora de preparar material didáctico. Las estrategias propuestas por diversos autores para enseñar a interpretarlos (Trimble 1985: 130-135 y 163-165) se basan, en esencia, en la intuición del hablante. Las relaciones semánticas internas implícitas de los sintagmas

[7] El mismo Lehrberger (1982: 92-94) no es preciso en la publicación donde aparecen las 50 relaciones y por tanto no se ha podido localizar. Sin embargo, la publicación debe ser de antes del año 1980.

[8] No existen obras sobre el tema en la literatura sobre la enseñanza del español para fines específicos.

nominales extensos especializados se explicitan mediante paráfrasis que, en muchos casos, no resuelven las supuestas ambigüedades de estos sintagmas nominales. El mismo Trimble (1985: 136) recomienda dejar los más extensos para el profesor del área de especialidad.

Quizá el estudio más importante cuantitativa y cualitativamente, desde la óptica del inglés para propósitos específicos (ESP), es el llevado a cabo por Salager Mayer (1984: 135-146). Esta autora compara los sintagmas nominales extensos del lenguaje general, el "lenguaje médico" y el "lenguaje de la técnica" llegando a la conclusión de que el promedio de la extensión entre los tres corpus es similar (2,06 en los tres). Sin embargo, el porcentaje de ocurrencia es más alto en medicina y en técnica (9,76% y 12,37%) que en el lenguaje general (0,87%)[9]. En cuanto a los sintagmas nominales extensos de cuatro y cinco palabras, la frecuencia en medicina y técnica es muy superior que en el lenguaje general (20 y 7 veces más, respectivamente). Esto significa que, si bien el promedio en la extensión es similar respecto del número de palabras totales, la distribución de la cantidad de tokens en los discursos especializados es mayor. Este estudio no hace ningún análisis de patrones gramaticales, relaciones semánticas, clases semánticas o distribución o función de los sintagmas nominales extensos en el discurso especializado.

2.8 Lingüística computacional e ingeniería lingüística

En el ámbito de la lingüística computacional y la ingeniería lingüística, algunos autores han trabajado las relaciones semánticas de los sintagmas nominales "cortos" teniendo en cuenta el concepto de '*compound*'[10]. Si un sintagma nominal extenso tiene, por ejemplo, seis elementos, se hacen tres

[9] De cada 100 palabras, menos de 1 palabra formaría un compuesto.

[10] Definido como la formación de una palabra a partir de más de dos palabras que funcionan como una unidad semántica.

pares de '*compounds*' y se explicitan las relaciones semánticas de los tres pares. Las relaciones o no entre los tres pares con el núcleo no se explicitan generalmente. Además, se han creado algunos '*braketers*' para identificar semiautomáticamente las relaciones semánticas de dos elementos (Baker y Szpakotwicz 1998: 96-102)

Muchos estudios (Woollie 1995; Thouvenin 1996) extraen las ocurrencias estadísticamente usando unos mínimos de frecuencia de 10 ó 25 apariciones y no usan etiquetaje automático o manual para observar las regularidades de los patrones. Thouvenin concluye que este método deja de lado muchos sintagmas nominales extensos de "baja frecuencia" que pueden ser términos en un área temática. Incluso advierte, desde la perspectiva del ESP, que estas unidades deberían aparecer en los diccionarios del área a pesar de su relativa fijación o baja frecuencia.

Por otro lado, es importante tener en cuenta que los etiquetadores automáticos pueden generar gran cantidad de ruido y silencio como se observa en el ejemplo 1.

1. [the trans-Golgi localized reversibly glycosylated polypeptide (RGP1)]
   xxx antibodies raised against [the **trans-Golgi localized** reversibly glycosylated polypeptide xxx] of
   MO6 NN6P H6 P [A666 **X V6A66** D6 X NN6S MO6] P
   [D X[11] V Adv X N X]

Puede verse que parte del sintagma nominal se interpretó como una forma verbal finita. Esto obstaculiza las búsquedas de manera automática o semiautomática. Si se piden por patrones generará mucho silencio como se ve en el ejemplo anterior.

[11] X: Sin identificar ni etiquetado.

Respecto del ruido, se realizó en el programa BwanaNet del *Corpus Tècnic* del IULA la siguiente consulta: N N N N N con mínimo 1-3 tokens obligatorios por cada N en todo el subcorpus de genoma.

Date:	Thu Jun 19 11:49:13 2003
Corpus:	Corpustecnicen
Subcorpus:	CORPUSTECNICEN:Last
Number of Matches:	103
Left display context:	7 tokens
Right display context:	7 tokens

Query text: CORPUSTECNICEN; a: [pos="N.*"]{1,3} [pos="N.*"]{1,3} [pos="N.*"]{1,3} [pos="N.*"]{1,3} [pos="N.*"]{1,3} :

Number of concordances: 103

1. Because of Watson-Crick affinity, the **Probes capture DNA strands hat** contain Boston's name (TCGGACTG).

NN6S Z</s> <s>P X NN6S Z A666 **NN6P NN6S NN6S NN6P NN6S** V6R6S NG6S NN6S Z X Z Z</s>

2. the latter approach, most often a **bait protein finds prey candidates**- sometimes many - and those might

A666 JA NN6S Z DS D6 A66S **NN6S NN6S NN6P NN6S NN6P** Z D6 R66A6666 Z C6 RD666666 V6666

3. and why we get old," **remarks geneticist Richard K. Wilson** of Washington University, one partner in

C6 D6 RE666N61P V6R6S JA Z Z **NN6P NN6S NN6S NN6S NN6S** P NN6S NN66 Z MC6 NN6S P

4. Most single mutations that increase **Activation produce amino acid substitutions** that increase the negative charge of region

NN6S Z</s> <s>A666 JA NN6P AD6S V6R6S **NN6S NN6S NN6S NN6S NN6P** X V6R6S A666 JA NN6S P NN6S

Puede verse que, en los casos anteriores, la desambiguación es errónea, pues el etiquetador analiza los verbos en tercera persona del singular o plural como sustantivos en plural o en singular, extrayendo sintagmas nominales extensos especializados erróneos. Por consecuencia, es difícil el reconocimiento y la extracción de estas unidades de tal extensión. Esto se debe quizá a que los etiquetadores se entrenaron con corpus de lengua general y no tienen en cuenta sintagmas de este tipo debido a su baja frecuencia respecto de los sintagmas de 1, 2 y 3 tokens.

2.9 Escritura técnica

Los autores de escritura técnica son –quizá- los principales detractores de los sintagmas con premodificación extensa y siempre advierten sobre el mal uso de ellos (Blake *et al* 1993; Burnett 1992; Huckin y Olsen 1991; Kirkman 1992; entre otros).

Por ejemplo, los autores de redacción técnica en inglés (*technical writing*) atacan contundentemente lo que llaman 'noun as adjective': "Strictly speaking, it is poor writing to use a noun as an adjective" (Blake y Bly 1993: 88).

Agregan que la premodificación extensa obscurece el significado. Como solución dicen que se debe romper esta premodificación con frases preposicionales o participiales como en los ejemplos 2 y 3.

2. a 15,000-lb steam/h pulverized-coal-fired power boiler

3. a pulverized-coal-fired power boiler generating 15,000 pounds of steam per hour

Otros autores ven los sintagmas nominales extensos como una característica que molesta al lector debido a que su premodificación es excesiva:

> "A particularly disturbing feature of technical writing is excessive 'premodification' – the piling up of adjectives, or words being used adjectivally, in front of a single noun: ... a mobile hopper fed compressed air operated grit blasting machine" (Kirkman 1992: 32-33).

Este mismo autor incluso compara la forma escrita con la oral. Por supuesto que en la forma oral no se usaría una premodificación tan compleja:

> "To pile up 'modifiers' in this way is utterly unnatural language behaviour. We would not normally dream of telling someone we had been to a store and bought a 'new green leather suede-lapelled patch-pocketed tie-belted jacket' " (Kirkman 1992: 32).

Burnett (1992: 312) analiza las cadenas de sustantivos como un problema estilístico: "Imprecise diction also results from noun strings: a series of two or more nouns in which the first nouns modify the later ones".

Además, sostiene que cuando una cadena de sustantivos alcanza las cinco o más palabras, se requiere tiempo extra para calcular las relaciones entre las palabras y, como consecuencia se puede interpretar de múltiples maneras y se vuelve indescifrable. La pregunta que cabe aquí sería para quién y en qué circunstancias una cadena es ambigua e indescifrable, y qué toma más tiempo y esfuerzo cognitivo para un lector experto, un sintagma nominal con múltiple posmodificación como sucede en las lenguas romances en las cuales se puede perder la referencia o un sintagma nominal con múltiple premodificación como sucede en inglés.

2.10 Comparación de algunos estudios

En la tabla 1 se han recogido los datos de varios estudios relacionados con los sintagmas nominales especializados extensos, terminológicos o libres (formados con unidades especializadas).

Autor	Corpus	Tokens	Método	2	3	4	5	6	7	8
Herzog 1971[12]	Computadores	ND[13]	ND	10%	36%	40%	12%	2%	-	-
Varantola 1984	Ingeniería/ noticias-general	20.000/ 11.526	Lingüístico	23,1/2 2,9	33,3/ 32,3	12,6/ 8,7	4,7/ 2,4	2,6/ 0,5	-	-
Salager 1984	Medicina/ técnica/ general	20.000/ 20.000/ 20.000	Estadístico	6,05/ 4,15/ 0,38	6,84/ 2,36/ 0,42	0,61/ 0,82/ 0,04	0,61 /0,2 2/,0 03	-	-	-
Goffin 1985[14]	ND	ND	ND	52	28	5+	-	-	-	
Horsella *et al* 1991[15]	Química	17.500[16]	ND	70,77 [17]	22,66	5,09	1,48	-	-	-
Montero 1995[18]	Computadores (diccionario)	4.235	Estadístico ND	82,9	14,89	1,9	0,1	-	-	-
Thouvenin 1996[19]	Electricidad	246.435	Estadístico	ND	ND	ND	ND	ND	-	-
Biber *et al* 1999[20]	Noticias/ prosa académica/ general	10.7 mill./ 5.3 mill./ 40 mill.	Lingüístico	80	20	2+		-	-	-
Café 1999	Biotecnología	ND	Lingüístico	53,8	37,61	7,61	0,96	-	-	-
Guzmán 2002[21]	Medicina	108.372	Lingüístico	44,50	27,87	14,57	8,51	2,44	2.4 2	
Oster 2003[22]	Cerámica	1.216	Lingüístico	57[23] 52,5	13,2 7,6	2,1 1,0	0,4 0,1	ND ND	0,1 0, 0	- -

[12] Se desconoce si es corpus textual, terminológico o lexicográfico.

[13] ND: No disponible.

[14] Citado por Cabré (1993). Estudio hecho para el alemán.

[15] Sólo trabaja con compuestos del tipo N+N+...N.

[16] Hace una comparación de tres niveles de especialización diferentes. Aquí sólo se han tomado los datos del nivel más especializado.

[17] Este porcentaje sólo se refiere al total de unidades pero no es frente al total de palabras.

[18] Este estudio no es sobre corpus de textos sino sobre el corpus de un diccionario.

[19] Se analizan las unidades de 2 a 4 tokens pero no hay estadística, patrones, ni frecuencias.

[20] Se analiza la premodificación múltiple, pero no hay estadísticas sobre el número de tokens de más de 3. Además, la estadística de los tokens de +3 se combina con posmodificación.

[21] En este estudio se cuentan como sintagmas las siglas en función de la cantidad de letras que forman la sigla como tokens independientes lo cual modifica enormemente los resultados y presenta tendencias no objetivas.

Quiroz 2003[24]	Medicina	44.000	Lingüístico	7,5	27	34	16,4	8,2	2,5	1,9
Cortés 2004[25]	Agrícola	17.7 mill.	Lingüístico	ND	ND	ND	ND	ND	ND	ND

Tabla 1: Cuadro comparativo de estudios sobre los SNEE

Estos estudios se han realizado en varias áreas del conocimiento (química, medicina, cerámica, etc.), varios niveles de especialidad (general vs. especializado), con diferentes métodos de análisis y extracción de datos (lingüístico, estadístico, manual, etc.). Puede observarse que en primer lugar, los métodos lingüísticos o híbridos dan porcentajes más altos, excepto Montero (1996) que toma un corpus de términos cerrado y no un corpus de textos.

En segundo lugar, hay diferencias muy grandes entre un autor y otro por las razones siguientes. El tipo de corpus empleado por cada autor tiene diferentes niveles de especialidad (además del concepto de nivel de especialidad que cada autor entiende), el número de tokens, el método de extracción y la combinación de subcorpus dentro de cada estudio (si los hay). Los corpus exclusivamente especializados presentan porcentajes más altos de aparición de este tipo de sintagmas y premodificación más extensa. Los corpus que combinan varios géneros o niveles de especialidad (desde el nivel de experto-lego hasta experto-experto) presentan porcentajes más bajos de sintagmas nominales especializados extensos y menos extensión de estos.

En tercer lugar, se observa que en casi todos los estudios se comienza con porcentajes más bajos en 2 premodificadores (sin sustantivo nuclear), se llega a un pico en los porcentajes con 3 y 4 premodificadores y, por último, se desciende desde 5 hasta 9 premodificadores, de modo que, existen menos

[22] El aspecto cuantitativo se realiza en un corpus terminológico alemán-español.
[23] Los datos de la primera línea se refieren al alemán y la segunda al español.
[24] Incluye determinantes al comienzo del sintagma.
[25] En este estudio no hay una estadística general sobre los números de tokens, los cuales hay que inferir de los datos del estudio. Por otro lado, se mezclan categorías léxicas abiertas y cerradas para la estadística, lo cual modifica los datos sensiblemente y presenta tendencias no objetivas.

sintagmas nominales con premodificación compleja entre más extensión haya en la premodificación. El hecho de haber un pico porcentual entre 3 y 4 puede revertir en estabilización y posiblemente en la lexicalización de estas unidades.

Por último, la mayoría de estos estudios no miran la premodificación de los sintagmas nominales extensos especializados desde la lexicalización, sino como un fenómeno que, en apariencia, se sale de los cánones de la lengua general para algunos y que para otros no se diferencia significativamente de la lengua general.

Para interpretar este cuadro es necesario saber qué observan algunos autores:

1. Cuántas palabras son parte de la premodificación respecto del total de palabras del corpus.

2. Cuántos sintagmas hay en un corpus y de estos cuántos corresponden a la premodificación múltiple.

3. Cuántos sintagmas nominales de premodificación compleja hay de ciertos tokens (¿de más de 2, 3 ó 4?). Luego se hace la estadística sobre ese parámetro de tokens y no sobre el total de sintagmas nominales de un corpus, es decir, desde un premodificador hasta el límite encontrado.

Algunos autores observan la opción 1 pero no las 2 y 3 o las opciones 2 ó 3 pero no la 1. Sólo un caso observa las opciones 1 y 2 (Salager-Mayer 1985). Otros autores observan la opción 3 y luego contabilizan el número total de tokens para sacar la ratio de palabras de los sintagmas nominales seleccionados con el total de palabras del corpus. Es cierto también que algunos estudios previos a los años 90 no disponían de corpus ni de herramientas para el procesamiento, el almacenamiento y la manipulación de los datos. A pesar de los avances en la creación de corpus y de las herramientas para su

procesamiento, los pocos estudios que hay en la actualidad tienen las mismas carencias que los anteriores.

Puede verse, entonces, que el asunto es mucho más complejo de lo que muestran las simples cifras. En primer lugar, no hay acuerdo sobre: a) los métodos de extracción de los sintagmas nominales, b) el tipo de corpus, c) el nivel de especialidad del corpus, d) el uso de lenguaje general versus lenguaje especializado, e) los métodos estadísticos para tratar y analizar los datos y, f) método de análisis lingüístico.

Además, la mayoría de los estudios, excepto Biber *et al* (1999) en algunos análisis, no tienen en cuenta: a) análisis de las categorías, b) análisis de patrones, c) análisis semántico (clases, relaciones, patrones, etc.), d) análisis morfológico, e) análisis tipográfico (números, cifras, siglas, etc.) y f) análisis textual (cohesión, etc.). El único estudio que tiene realmente un análisis semántico es el de Oster (2003) para el par alemán-español.

Como consecuencia, no hay una visión amplia ni un conjunto amplio de análisis que permita hacer generalizaciones sobre este fenómeno. Una visión parcial de este fenómeno se ve favorecida por:

1. La baja frecuencia de los sintagmas nominales extensos especializados en el conjunto de la lengua general.

2. Algunos estudios sólo se centran en el discurso especializado y en unos niveles de especialidad que en muchos casos pueden presentar una cantidad baja de sintagmas nominales extensos especializados y se encuentran muy poca cantidad del tipo de texto que aquí se estudia (‘*journal*’).

3. En general, los corpus de los diferentes estudios, excepto Biber *et al*, no podrían considerarse como representativos tanto en lenguaje general como especializado, o ambos, debido al tamaño de los corpus (muy pequeños), los tipos de texto, etc. Aún el corpus de base de Biber *et al* puede ser discutible en

cuanto a que se hacen generalizaciones sobre el discurso especializado (llamado '*academic prose*').

Un aspecto importante es que casi todos los estudios provienen de investigadores de ESP ('*English for Specific Purposes*') o traducción, puesto que es en estas disciplinas donde este fenómeno se considera un verdadero problema.

Sin embargo, los mayores aportes los han hecho los gramáticos (Quirk *et al* 1985; Biber *et al* 1999; Huddleston y Pullum 2002) en aspectos como a) la categoría gramatical que predomina en la premodificación, b) el orden de los premodificadores, las restricciones categoriales, el orden "natural", etc. y, c) las relaciones y clases semánticas que pueden ocurrir en la premodificación. Esto no se especifica para los sintagmas con premodificación larga o muy larga, pero pueden ser muy útiles para nuestro estudio, es decir, nos puede ayudar a observar el orden de los premodificadores, las relaciones y clases semánticas que predominan en el discurso especializado y las funciones que cumplen.

3. Metodología general de trabajo y constitución de corpus

3.1 Introducción

Para alcanzar los objetivos propuestos y demostrar las hipótesis planteadas, se configuró el siguiente corpus y se llevó a cabo la siguiente metodología y análisis de datos.

3.1.1 Descripción y selección de los corpus de referencia

Para esta tesis se han empleado tres tipos de corpus de referencia para la extracción de los diferentes subcorpus de análisis: un corpus en inglés, un corpus en español y un corpus paralelo inglés-español. Adicionalmente, se ha empleado un corpus lexicográfico compuesto por cinco diccionarios en formato electrónico como corpus de contraste.

Durante años, el autor de la presente tesis trabajó no sólo como traductor de ciencias de la salud: medicina, enfermería, veterinaria, biología sino como profesor de traducción e investigador de traducción y terminología en dichas áreas. De igual modo, durante el período de la tesis, ya existía un corpus considerable de medicina en diferentes niveles de especialidad en el *Corpus Tècnic* del IULA. Además, se estaba confeccionando un subcorpus de genoma en el marco del proyecto Genoma. Por estas razones, se decidió seleccionar la medicina y el área de genoma como el ámbito de especialización en el que enmarcaríamos nuestro objeto de estudio.

Para poder tener controlados todos los datos en cuanto a la variación horizontal, se decidió no emplear un corpus con varias áreas o ámbitos del conocimiento. El tener varios ámbitos, abriría otras puertas pero también variables que no podríamos controlar de manera fiable dada la configuración del corpus que nos habíamos planteado desde un comienzo. De todos modos, y

como una manera de reforzar nuestra hipótesis sobre la existencia y la frecuencia de este fenómeno en la lengua, se ha empleado un corpus lexicográfico de otras áreas del conocimiento: estadística, economía, finanzas y medicina.

Controlar la homogeneidad temática del corpus textual fue complicado a la hora de conseguir un corpus paralelo de la misma temática puesto que no es fácil encontrar revistas traducidas del inglés al español del nivel experto a experto. En las pocas revistas que existen, no era fácil delimitar un artículo que perteneciera al genoma. Por tanto, se decidió compilar *ad hoc* el corpus paralelo de la revista *The Lancet*, no sólo por su prestigio en el área de la medicina sino porque se produjo una versión española hasta 1999. Los artículos se seleccionaron principalmente teniendo en cuenta el formato IMMRAD (*Introduction, Materials and Methods, Results and Discussion*) y su disponibilidad en la versión en papel del español. Todos los textos comprenden el período entre 1997 y 1998 y se han procesado en el *Corpus Tècnic* del IULA.

Entre los criterios del corpus de genoma del IULA es importante resaltar las áreas involucradas:

Farmacogenómica
Neurociencia
Enfermedades
Eugenesia
Biotecnología
Diferenciación
Inmunología
Investigación genética
Estructura interna
Ingeniería genética
Filogenia

La selección de esta combinación de corpus, obedece a la combinación de varios factores.

En primer lugar, se observarán cuantitativa y cualitativamente los patrones más frecuentes tanto sintácticos como semánticos en un corpus especializado en inglés.

En segundo lugar, se observarán cuantitativa y cualitativamente los patrones más frecuentes tanto sintácticos como semánticos en un corpus especializado en español.

En tercer lugar, se estudiarán cuantitativa y cualitativamente los patrones sintácticos más frecuentes en inglés y sus respectivas soluciones en español en el corpus paralelo del inglés al español. Se contrastarán con los resultados obtenidos en los corpus anteriores para observar, en primer lugar, los patrones más frecuentes en inglés y sus respectivas soluciones en español y, en segundo lugar, observar si las estructuras antes descritas en inglés y español siguen la tendencia mostrada o no, de modo que permitan observar si hay interferencias en los traductores en cuanto a las soluciones de traducción al español por parte del inglés.

En cuarto lugar, se contrastará si las tendencias de los corpus de análisis en cuanto a la extensión y los patrones sintácticos están presentes en el corpus lexicográfico no sólo de medicina sino de las otras áreas. De este modo, puede confirmarse o no las tendencias presentadas en algunos estudios hechos en corpus lexicográficos (Montero 1995) y observar si la realidad traductiva se ve reflejada en las fuentes de consulta que son los diccionarios o por el contrario los traductores deben emplear otros procedimientos para llegar a la solución de este tipo de sintagmas.

Igualmente, se analizará si las tendencias observadas en los corpus de análisis del inglés y el español extraído de los corpus de referencia se manifiestan en el corpus lexicográfico.

En último lugar, se contrastarán los resultados obtenidos tanto en el corpus de análisis del español como en el corpus paralelo con los datos obtenidos en el corpus CREA de la RAE. Como corpus general del español, podrá observarse si las tendencias cuantitativas de los corpus especializados se reflejan en un corpus de lengua general como el CREA o por el contrario, divergen en la frecuencia de aparición de las estructuras.

Esta combinación de corpus y análisis sirve para contrastar los análisis hechos por otros autores que sólo han trabajado con un tipo de corpus o que no han contrastado los resultados con otros corpus como se expuso en §2.10. Por tanto, los resultados obtenidos en esta tesis serán más confiables y generalizables no sólo para traductores sino para otros profesionales o investigadores en las ciencias del lenguaje, lexicógrafos, terminólogos y profesores de ESP y traducción.

3.1.2 Descripción y selección del corpus en inglés

Se seleccionó un corpus de 128 textos en inglés con aproximadamente 476.337[26] palabras a partir de los 257 textos (1.303.576 palabras) del *Corpus Técnic* del IULA. Todos los textos escogidos se tomaron de varias revistas, entre ellas *The Lancet, Genomics* y *FEBS Letters* con el formato IMMRAD (*Introduction, Materials and Methods, Results and Discussion*). Los criterios de selección para cada texto son:

- Pertenecer al área del genoma
- Estar escrito por un hablante nativo del inglés: se observó que al menos uno de los autores tuviera apellidos de origen inglés. Si esto no se podía establecer, entonces se tuvo en cuenta:

[26] Datos procedentes del *Corpus Tècnic* del IULA de la UPF (CT-IULA) obtenidos a través de BwanaNet en noviembre de 2004.

- que estuviera escrito en un país de habla inglesa (Reino Unido, Estados Unidos, Canadá y Australia, principalmente),
- que por lo menos un laboratorio o universidad de habla inglesa estuviera involucrado en la redacción.

- Estar disponible en versión electrónica

3.1.3 Descripción y selección del corpus en español

Se seleccionó un corpus de aproximadamente 86 textos que equivalen a 464.333 palabras tomado de los 278 textos (1.693.515 palabras) del corpus de Genoma del *Corpus Tècnic* del IULA. Los textos pertenecen tanto a revistas con el formato IMMRAD como a capítulos de libros y tesis doctorales. Los criterios de selección para cada texto son:

- Pertenecer al área del genoma
- Estar escrito por un hablante nativo del español: se observó que al menos uno de los autores tuviera apellidos de origen hispano. Si esto no se podía establecer, entonces se tuvo en cuenta:
  - que estuviera escrito en un país de habla española (Latinoamérica y España),
  - que por lo menos un laboratorio o universidad de habla española estuviera involucrado en la redacción.
- Estar disponible en versión electrónica

En resumen, la tabla 1 muestra los datos empleados para confeccionar los corpus de referencia y el corpus paralelo a partir del corpus general del IULA.

	Inglés	Español
Corpus Técnic del IULA	1.303.576	1.693.515
N.º de textos totales IULA	257	278
Corpus seleccionado	476.337	464.333
N.º de textos seleccionados	128	86
Corpus paralelo	66.534	86.457
N.º de textos totales	21	21

Tabla 1: Número de textos y palabras de los corpus.

3.1.4 Descripción y selección del corpus paralelo inglés-español

Para observar las regularidades de los patrones y de las soluciones en la traducción de la premodificación del inglés al español, se recogió un corpus de 66.534 palabras a partir de 21 textos en inglés. Todos los textos son artículos de investigación con la estructura IMMRAD de la revista médica *The Lancet*[27]. Esta revista se tradujo completamente en español hasta 1999 y por tanto, la selección de los textos se hizo de 1997 a 1998. Todas las secciones fueron guardadas excepto el resumen, los nombres y la afiliación institucional del autor, los agradecimientos, y las referencias bibliográficas. Debido a problemas técnicos, algunos gráficos y tablas con texto relevante tuvieron que ser eliminados. En general, los textos se procesaron según las indicaciones del *Corpus Tècnic* del IULA. El número promedio de palabras por texto en el corpus es de 3.168 con un mínimo de 2.028 palabras y un máximo de 4.783.

De acuerdo con la versión española de *The Lancet*, los artículos fueron traducidos por reconocidos expertos en medicina: profesores e investigadores.

[27] Los textos en inglés se recogieron de los volúmenes y de las ediciones siguientes: 349 (marzo de 1997), 351 (enero, febrero y marzo de 1998) y 352 (octubre de 1998).

3.1.5 Descripción y selección del corpus lexicográfico

Para verificar la existencia y la frecuencia de este fenómeno en otras áreas del conocimiento y por extensión a la lengua en general, así como los patrones más frecuentes en los recursos terminológicos, se ha constituido un corpus de cinco diccionarios electrónicos disponibles en CD-ROM, Word, PDF o HTML: Diccionario Mosby de medicina, el Diccionario inglés-español de Ciencias de Laboratorio Clínico -IFCC, *IMF Terminology*, *Routledge Spanish Dictionary of Business, Commerce and Finance* e *ISI Multilingual Glossary of Statistical Terms*. Estos diccionarios pertenecen a diferentes áreas del conocimiento y varían de tamaño. De cada uno de ellos, sólo se usaron los términos en inglés y en español de 3 o más tokens de categoría gramatical abierta (sustantivo, adjetivo, adverbio y verbo). No se seleccionaron las unidades con posesivo sajón, unidades coordinadas (*and*, *or*, y o) o posmodificadas en inglés. A continuación, se describe brevemente cada diccionario.

Diccionario	Área temática	N.º de entradas	SN de +3 tokens en inglés	Porcentaje	SN de +3 tokens en español	Porcentaje
Diccionario Mosby	Medicina	31.400	3.553	11,31%	3.848	12,25%
Diccionario IFCC	Lab. clínico	4.039	725	17,94%	608	15,05%
IMF Terminology	Economía	4.500	766	17,02%	1.367	30,37%
Routledge Dictionary	Finanzas	38.000	5.269	13,86%	1.491	3,92%
ISI Multilingual Glossary	Estadística	3.500	1.238	35,37%	921	26,31%

Tabla 2: Resumen de datos del corpus lexicográfico.

1. Diccionario Mosby de medicina, enfermería y ciencias de la salud (2000), 5ta edición inglés-español: diccionario en formato chm (archivo tipo ayuda). Este diccionario es la versión en lengua española de la 5. ª edición de la obra original en inglés: *Mosby's Medical, Nursing, and Allied Health Dictionary*. Contiene unas 31.400 entradas en ambas lenguas de las cuales 3.553 entradas en inglés tienen más de 3 tokens (11,31%) y 3.848 entradas en

español tienen más de 3 tokens (12,25%). Además, contiene definiciones en español, referencias cruzadas, gráficos y entre otros campos. Es un diccionario dirigido a los profesionales de las ciencias de la salud.

2. Diccionario inglés-español de Ciencias de Laboratorio Clínico IFCC[28]: Glosario inglés-español del Grupo de Trabajo sobre Terminología y Nomenclatura en Química Clínica en Lengua Española de la Federación Internacional de Química Clínica - División Científica. Aunque no se expresa explícitamente, es un glosario normativo y consensuado para todos los países de habla hispana. Está actualizado hasta el año 2000. Contiene 4.039 entradas en ambas lenguas de las cuales 725 entradas en inglés tienen más de 3 tokens (17,94 %) y 608 entradas en español tienen más de 3 tokens (15,05%).

3. *IMF Terminology*[29]: La base de datos de terminología del Fondo Monetario Internacional contiene 4.500 registros en inglés, español, alemán, portugués y francés sobre finanzas y economía. Esta base de datos incluye sólo equivalentes en cada lengua sin definiciones. También incluye frases, nombre de instituciones, acrónimos, referencias cruzadas, contextos, etc. Es una base de datos dirigida especialmente a traductores. Está actualizada hasta el año 2000. Contiene 4.500 entradas en inglés y en español, de las cuales 766 entradas en inglés tienen más de 3 tokens (17,02 %) y 1.367 entradas en español tienen más de 3 tokens (30,37%).

4. *Routledge Spanish Dictionary of Business, Commerce and Finance*/ Diccionario Inglés de Negocios, Comercio y Finanzas (1999). Contiene más de 38.000 términos en inglés y español de negocios, comercio y finanzas en 45 subáreas relacionadas. Este diccionario se confeccionó con base en la versión impresa del Diccionario Inglés de Comercio, Negocios y Finanzas Routledge de 1998. Contiene unas 38.000 entradas en ambas lenguas de las cuales 5.269

[28] Disponible en http://www.leeds.ac.uk/ifcc/PD/dict/spandict.html.

[29] Disponible en http://www.imf.org/external/np/term/index.asp?index=eng&index_langid=1.

entradas en inglés tienen 3 tokens (13,86 %) y 1.491 entradas en español tienen 3 tokens (3,92%).

5. *ISI Multilingual Glossary of Statistical Terms*[30]: Diccionario del *International Statistical Institute* con más de 3.500 términos sobre estadística y áreas relacionadas en 21 lenguas. El glosario va dirigido especialmente a expertos en el área. Se actualiza constantemente, tanto el número de entradas como número de lenguas. Contiene 3.500 entradas en ambas lenguas de las cuales 1.238 entradas en inglés tienen 3 tokens (35,37 %) y 921 entradas en español tienen 3 tokens (26,31%).

3.2 Herramientas y recursos

A lo largo de todo el trabajo de investigación, se emplearon diversos tipos de herramientas para procesar, almacenar y presentar los datos, excepto para el corpus de referencia del inglés y el español que ya estaban procesados con las herramientas del *Corpus Tècnic* del IULA como parte del Banco de conocimiento GENOMA.[31]

El banco *Genoma* que contiene información textual, documental, terminológica y conceptual sobre genómica humana es una herramienta de consulta para traductores, redactores y especialistas en la materia.

[30] Disponible en http://isi.cbs.nl/glossary/index.htm.

[31] Este banco fue desarrollado por el grupo IULATERM en el marco de los proyectos TEXTERM (BFF 2000-0841) y RICOTERM (TIC 2000-1191). Toda la información está disponible en http://genoma.iula.upf.edu:8080/genoma/index.jsp.

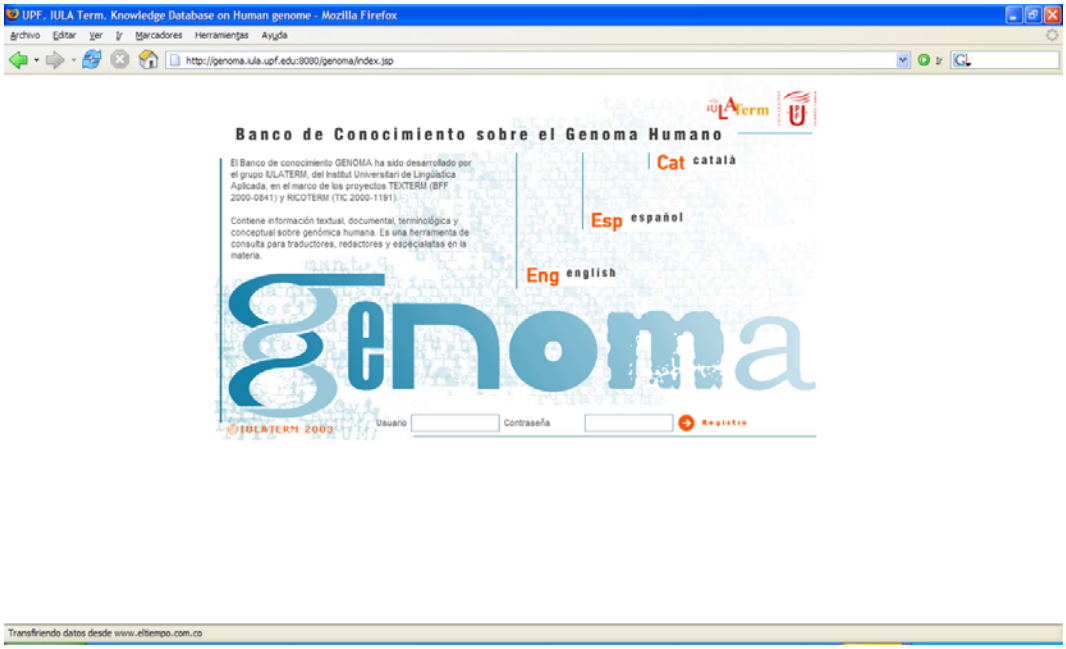


Figura 1: Banco de conocimiento sobre el Genoma Humano del IULA.

Para construir los primeros patrones y probar su existencia en un corpus piloto, se empleó el programa Repóker del IULA (Quiroz *et al* 2004).

La extracción de los sintagmas de los corpus de referencia del *Corpus Tècnic* del IULA se realizó mediante una serie de herramientas y *scripts* en Perl del programa Bwananet.

Para conseguir el número de tokens de cada sintagma, procesar los diferentes diccionarios del corpus lexicográfico, los corpus de análisis y paralelo y realizar diversas tareas de procesamiento de datos se emplearon varios *scripts* realizados para el programa Perl.

El almacenamiento de todos los datos de referencia, muestras, listas de los diccionarios, cálculos de algunas estadísticas y manipulación de todos los datos se hizo en el programa de hoja de cálculo Excel 2003 de Microsoft.

Igualmente, se empleó el programa de edición de texto Editplus para procesar los datos y poder elaborar las diferentes listas de análisis.

Para el procesamiento de la estadística descriptiva y la creación de tablas e informes derivados de ella, se empleó el programa de estadística Statgraphics Pro 5.1.

El marcaje semántico de la muestra se hizo con el programa WordNet 2.1 de la Universidad de Princeton para el inglés y para el español, se empleó la versión europea de WordNet, EuroWordNet 1.6. En inglés, también se empleó el conjunto de recursos UMLS versión 2006AC.

Para corroborar datos, principalmente durante el etiquetaje semántico, se emplearon varios diccionarios de referencia tanto generales como de medicina, en especial, los diccionarios en CD-ROM: Diccionario Mosby de medicina 2000, *Stedman's Medical Dictionary 3.0*, Diccionario Espasa de Medicina, Diccionario de la Real Academia Española, Diccionario Webster en inglés, Diccionario Vox de la lengua española y el *Collins English Dictionary*.

3.2.1 Herramientas de etiquetaje

Para poder extraer los patrones superficiales y obtener las regularidades semánticas, los datos se han etiquetado con varias herramientas. Para etiquetar los datos del corpus lexicográfico, se empleó *Machinese Phrase Tagger* online demo y para etiquetar semánticamente las muestras de los patrones se utilizaron los programas WordNet 2.1 y UMLS 2006 AB y AC en línea para el inglés y EuroWordNet 1.6 en línea para el español. A continuación se describe brevemente cada uno de los programas.

3.2.1.1 Machinese Phrase Tagger online demo

Para etiquetar el corpus lexicográfico, se usó el etiquetador *Machinese Phrase Tagger online demo*. *Machinese Phrase Tagger* es un programa que realiza tareas básicas de análisis lingüístico y proporciona la información

relevante sobre las palabras a cantidades grandes de texto. *Machinese Phrase Tagger* divide el texto en unidades de palabra y le asigna etiquetas morfosintácticas a cada una. Los desarrolladores son los mismos de la *Constraint Grammar* que ahora crean nuevas herramientas lingüísticas para el procesamiento del lenguaje natural en varias lenguas.

3.2.2.2 WordNet 2.1

Para el etiquetaje semántico en inglés, se utilizó el programa WordNet 2.1 de la Universidad de Princeton[32].

WordNet es la base de datos léxica más grande en lengua inglesa, desarrollada bajo la dirección del Prof. George Miller. Las categorías léxicas abiertas como sustantivos, verbos, adjetivos y adverbios se agrupan en los sistemas de los sinónimos cognoscitivos llamados *synsets* (*synonym sets*), en el que cada uno expresa un concepto distinto.

Un *synset* es un sistema de palabras con la misma categoría léxica que puede intercambiarse en determinados contextos. En el siguiente ejemplo, extraído de EWM 1.5, el conjunto de palabras {carro, coche, automóvil, auto, máquina} es un *synset* porque pueden ser utilizadas para referir al mismo concepto. Este *synset* puede describirse como: "un aparato de 4 ruedas, propulsado generalmente por un motor de combustión interna". Finalmente, los *synsets* pueden relacionarse los unos con los otros mediante relaciones semánticas, tales como hiperonimia/hiponimia, superordinado/subordinado, antonimia, implicaciones y meronimia/holonimia, como se ilustra en la figura 2.

[32] Este programa se descargó bajo licencia de la Universidad de Princeton en http://wordnet.princeton.edu/.

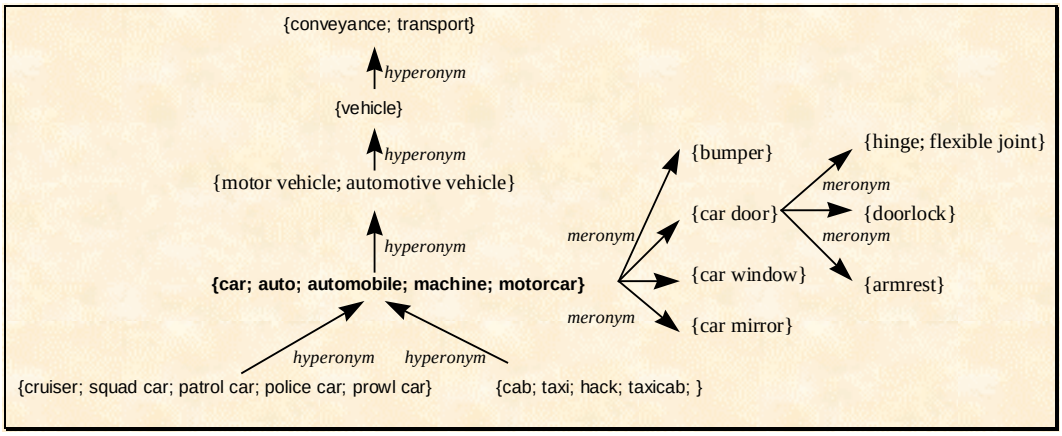


Figura 2: Ejemplo de la jerarquía de las relaciones en un *synset*.

Una palabra o una colocación puede aparecer en más de un *synset* y en más de una categoría gramatical.

Los sustantivos y los verbos se organizan en las jerarquías con base en las relaciones de hiperonimia/hiponimia entre los *synsets*. En cambio, los adjetivos están organizados en "clusters" que contienen los *synsets* principales y los *synsets* con base en los satélites. Cada "cluster" se organiza alrededor de pares antónimos (y tríos en algún caso). Los pares antónimos (o tríos) se indican en los *synsets* principales de un "cluster". La mayoría de *synsets* núcleo tienen unos o más *synsets* con base en los satélites, que representan al concepto, el cual es similar en el significado al concepto que representa al *synset* principal (WordNet 2005).

Los "pertainyms" son adjetivos relacionales y no siguen la estructura antes descrita. Los "pertainyms" no tienen antónimos; el *synset* para el "pertainym" contiene solo una palabra o colocación y un indicador léxico al sustantivo del cual se deriva el adjetivo. Los adjetivos participios tienen indicadores léxicos a los verbos de los cuales se derivan.

Los adverbios se derivan generalmente de los adjetivos y tienen antónimos en algunos casos. Por tanto, el *synset* para un adverbio contiene generalmente el indicador léxico del adjetivo del cual se deriva.

A continuación se presenta la lista de las 25 clases de sustantivos del nivel superior de WordNet denominadas “Tops”:

act, action, activity
animal, fauna
artifact
attribute, property
body, corpus
cognition, knowledge
communication
event, happening
feeling, emotion
food
group, collection
location, place
motive
natural object
natural phenomenon
person, human being
plant, flora
possession
process
quantity, amount
relation
shape
state, condition
substance
time

En la tabla 3 se presenta un resumen de los datos generales de la versión 2.1 de WordNet.

Categoría léxica	Cadenas únicas	*Synsets*	Total pares palabra-sentido
Sustantivo	117.097	81.426	145.104
Verbo	11.488	13.650	24.890
Adjetivo	22.141	18.877	31.302
Adverbio	4.601	3.644	5.720
Totales	155.327	117.597	207.016

Tabla 3: Número de palabras, *synsets* y sentidos de WordNet 2.1.

Huelga decir que la estructura y datos de WordNet le hacen una herramienta muy útil para la lingüística de computacional y el procesamiento de lenguaje natural.

3.2.2.3 EuroWordNet 1.6 para el español

Para etiquetar los datos del español, se empleó EuroWordNet[33] (EWN) en línea[34]. EuroWordNet es una base de datos léxica multilingüe con los WordNets para varias lenguas europeas, entre ellas el español, siguiendo las mismas líneas que el WordNet de la Universidad de Princeton (Fellbaum 1998). WordNet contiene información sobre sustantivos, verbos, adjetivos y adverbios en inglés y se organiza alrededor de la noción de un *synset* como se explicó en el apartado anterior.

A pesar de que EuroWordNet tiene como base la estructura de WordNet 1.5, la idea de *synset* y las relaciones semánticas principales, se hicieron algunos cambios en la base de datos, de modo que reflejara:

La idea de una base de datos multilingüe
Las relaciones específicas de cada lengua
La máxima compatibilidad entre los diferentes recursos

[33] EuroWordNet fue un proyecto financiado por la Unión Europea. El proyecto comenzó en marzo de 1996 y terminó en 1999 en su primera fase.

[34] Disponible en http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl o http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl.

La construcción de Wordnets relativamente independientes (re)utilizando los recursos existentes.

La diferencia más importante de EuroWordNet con respecto a WordNet es su carácter multilingüe, que sin embargo también plantea algunas preguntas fundamentales con respecto al estado de la información monolingüe en los WordNets como los han planteado los propios creadores (Vossen 1999: 8). En principio, el carácter multilingüe se logra agregando una relación de equivalencia para cada *synset* en una lengua al *synset* más próximo de WordNet 1.5. Los *synsets* ligados al mismo *synset* de WordNet 1.5 se supone que son equivalentes o tienen un significado cercano y pueden entonces ser comparados. La diferencia radica en que si las palabras equivalentes se relacionan de diversas maneras en los diversos recursos y, por tanto, se debe validar dicha diferencia. Como la misma documentación lo manifiesta, en el WordNet en holandés se puede observar que *hond* (perro) está clasificado tanto como *huisdier* (animal doméstico) como *zoogdier* (mamífero). Sin embargo, no hay equivalente para *pet* (animal doméstico) en italiano, y *cane* del italiano, que está relacionado con el synset *dog* (perro), se clasifica solamente como *mammal* (mamífero) en el Wordnet en italiano (Vossen 1999: 8).

A continuación se presenta la lista de las 59 categorías de sustantivos del nivel superior de EuroWordNet para el español[35]:

Vehicle
SituationType
Container
Place
Phenomenal
Comestible
Static

[35] Para más información sobre el significado de estas categorías, puede consultarse el sitio de EuroWordNet para el español: http://siuc01.si.ehu.es/cgi-bin/mcr/public/wei.definitions.perl#vehicle.

Existence
Software
Garment
Building
Functional
ImageRepresentation
Communication
Part
Object
LanguageRepresentation
Instrument
Physical
Covering
Relation
Quantity
Manner
Mental
3rdOrderEntity
BoundedEvent
Furniture
Property
Dynamic
UnboundedEvent
Function
Condition
Substance
Experience
Liquid
Living
Group
Modal
Purpose
Artifact
Time
Stimulating
Cause

1stOrderEntity
Animal
Representation
Agentive
Usage
Occupation
Possession
Natural
Human
Location
Solid
Social
Creature
Gas
MoneyRepresentation
Plant

En la tabla 4, se muestran los datos de la base de datos de EWN para el español.

Categoría léxica	Total de sentidos (variantes)	*Synsets*	Total pares palabra-sentido
Sustantivo	40.759	24.215	26.485
Verbo	9.317	4.079	3.828
Otros	2.439	2.191	2.439
Totales	52.515	30.485	32.752

Tabla 4: Número de palabras, *synsets* y sentidos de EWN 1.6 para el español.

A continuación, en la figura 3 puede ver la interfaz de WordNet 1.6 para el español con el ejemplo “enzima”.

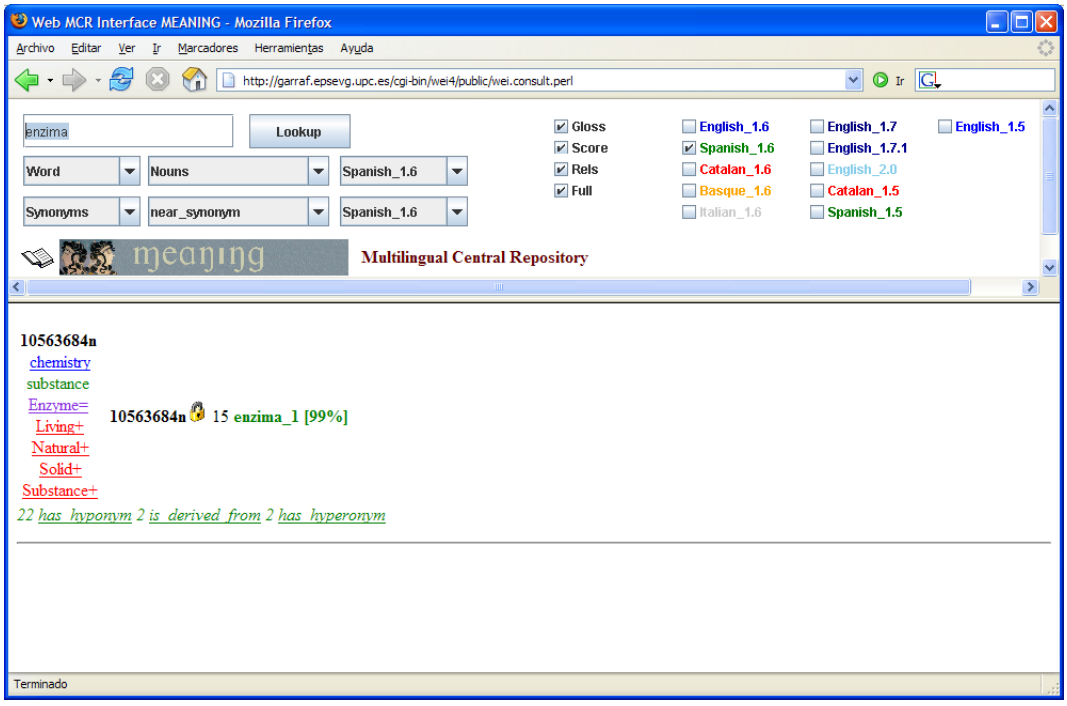


Figura 3: Interfaz de consulta de Wordnet 1.6 en español.

3.2.1.4 UMLS 2006 AB/AC

Dado que WordNet 2.1 es una ontología para propósitos generales y que el tema de tesis se enmarca en la medicina, se decidió etiquetar también la muestra del inglés con el programa UMLS 2006AB/AC[36] en línea, *UMLS Knowledge Source Server* (UMLSKS)[37].

UMLS (*Unified Medical Language System*), creado y diseñado por la *National Library of Medicine (NLM)*, es un conjunto de recursos léxicos que se crearon con el propósito de hacer legibles los datos médicos para los computadores.

El propósito de UMLS es facilitar el desarrollo de sistemas informáticos que "entiendan" el lenguaje de la biomedicina y la salud. Los datos de UMLS

[36] Esta base de datos se utilizó bajo licencia de la National Library of Medicine (NLM).
[37] Los recursos y las herramientas de UMLS Knowledge Source Server (UMLSKS) se pueden acceder en http://umlsks.nlm.nih.gov, bajo previa licencia pedida por el usuario.

son multiusos y pueden usarse en sistemas que emplean diferentes tipos de información médica como historias clínicas, literatura científica, normas, datos de salud pública y administración de bibliotecas médicas.

UMLSKS[38] es un conjunto de herramientas web que permite al usuario y a programadores acceder a las terminologías biomédicas de UMLS. En este portal web, se encuentran los tres repositorios de datos de UMLS.

UMLS Metathesaurus: contiene la información sobre conceptos y términos biomédicos de más de 100 vocabularios y clasificaciones controlados que se emplean en historias clínicas, datos administrativos de salud, bases de datos bibliográficos y de texto y sistemas expertos.

Semantic Network: a través de sus tipos semánticos, la red semántica proporciona una categorización consistente en todos los conceptos que están representados en el metatesauro de *UMLS*. Los enlaces entre los tipos semánticos proporcionan la estructura para la red semántica y representan relaciones importantes en el ámbito biomédico.

SPECIALIST Lexicon: lexicón en inglés con términos de biomedicina que contiene información sintáctica, morfológica, y ortográfica para cada término o palabra.

Los vocabularios fuente del *UMLS Metathesaurus* incluyen terminologías diseñadas para ser empleadas en sistemas de historias clínicas y grandes clasificaciones de procedimientos y enfermedades que se utilizan para preparar informes estadísticos y facturas. Los vocabularios más específicos se usan para guardar datos relacionados con la psiquiatría, la enfermería, los aparatos médicos, las reacciones secundarias de las drogas, etc. Igualmente, las terminologías de UMLS sobre enfermedades y hallazgos médicos se emplean en

[38] Para más información, se puede acceder a la documentación en línea de UMLSKS en http://www.nlm.nih.gov/research/umls/documentation.html.

sistemas expertos de diagnóstico y algunos tesauros se usan para la recuperación de información. También existen una lista categorizada de los vocabularios fuente del inglés de más de 100 terminologías, clasificaciones y tesauros, algunos en ediciones múltiples.

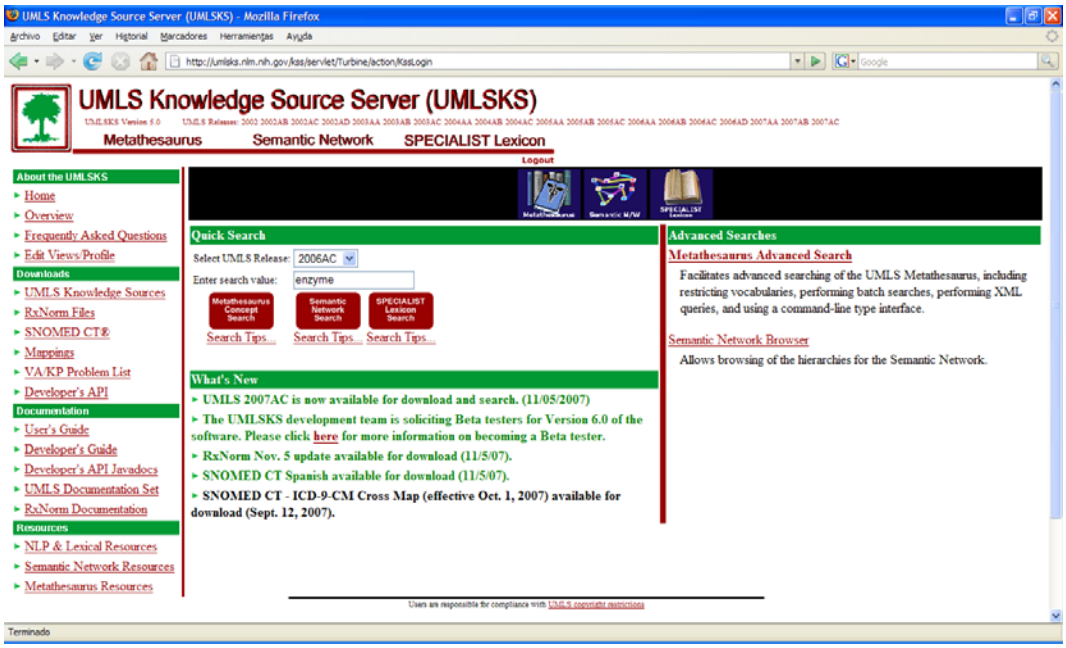


Figura 4: Interfaz de consulta de UMLSKS en inglés.

En la figura 4, se muestra la interfaz de consulta de UMLSKS. Obsérvese que los recursos se encuentran ubicados en el marco izquierdo. En el marco del centro aparece el cuadro de búsqueda y las tres opciones con los diferentes recursos antes descritos.

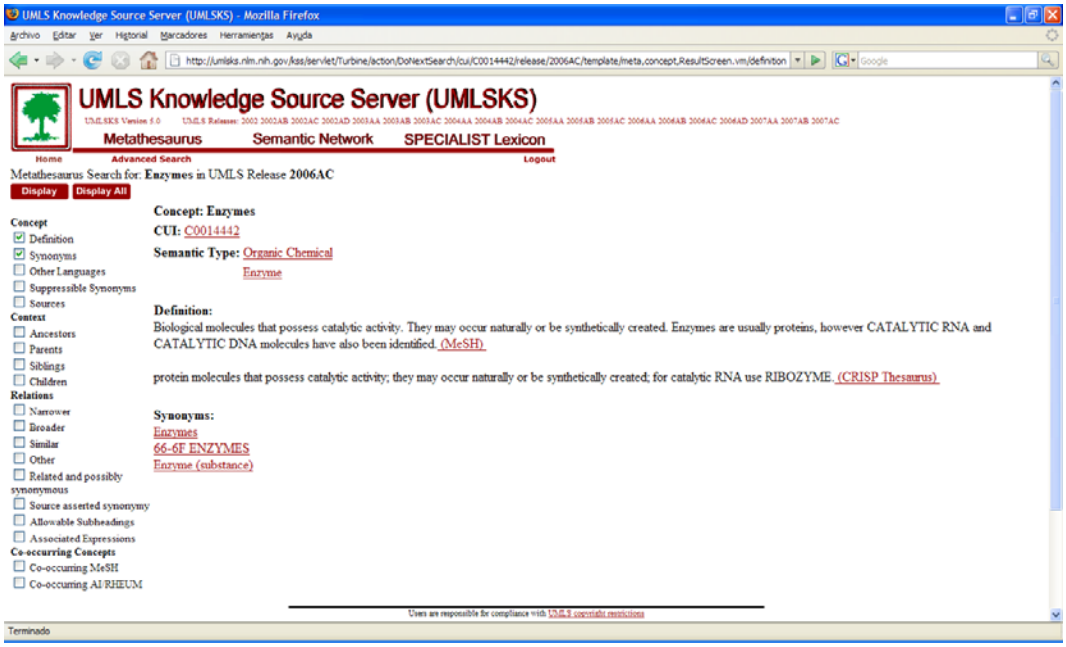


Figura 5: Resultados de una consulta en UMSLKS.

En la figura 5, puede verse el resultado de una búsqueda en UMLSKS, concretamente el término *enzyme*. En primer lugar, se despliega, el nombre del concepto, su número de identificación, los tipos semánticos a los que pertenece el término y que se han empleado en este trabajo, las definiciones y sus fuentes, los sinónimos y sus tipos semánticos entre paréntesis, entre otros campos.

Algunos datos importantes sobre UMLS para la versión 2006AC son:

Número de conceptos: 1.371.699
Número de nombre de conceptos: 6.499.598
Número de nombre de conceptos en inglés: 4.284.888
Número de nombre de conceptos diferentes: 5.369.057
Número de nombre de conceptos diferentes normalizados: 4.789.290
Número de fuentes (familias de fuentes diferentes por idioma): 120
Número de fuentes que contribuyen con nombre de conceptos: 138
Número de idiomas que contribuyen con nombre de conceptos: 17

UMLS tiene una red semántica denominada “Semantic Network”, como se observa en la figura 5. Dicha red semántica tiene el propósito de reducir la

complejidad del metatesauro, agrupando los conceptos de acuerdo con los tipos semánticos que se les han asignado. Sin embargo, para poder obtener mejores generalizaciones como se presenta en §6.4.5 y ss. es preferible un conjunto más pequeño y granulado de tipos semánticos. Por eso, la red semántica cuenta con 15 grupos semánticos que proporcionan una partición del *UMLS Metathesaurus* para el 99,5% de los conceptos.

La red semántica de UMLS contiene actualmente 134 tipos semánticos y 54 relaciones semánticas. La red se define en su nivel más alto en dos jerarquías, una para las entidades "Entity" y otras para los eventos "Events", como se muestra en la figura 6.

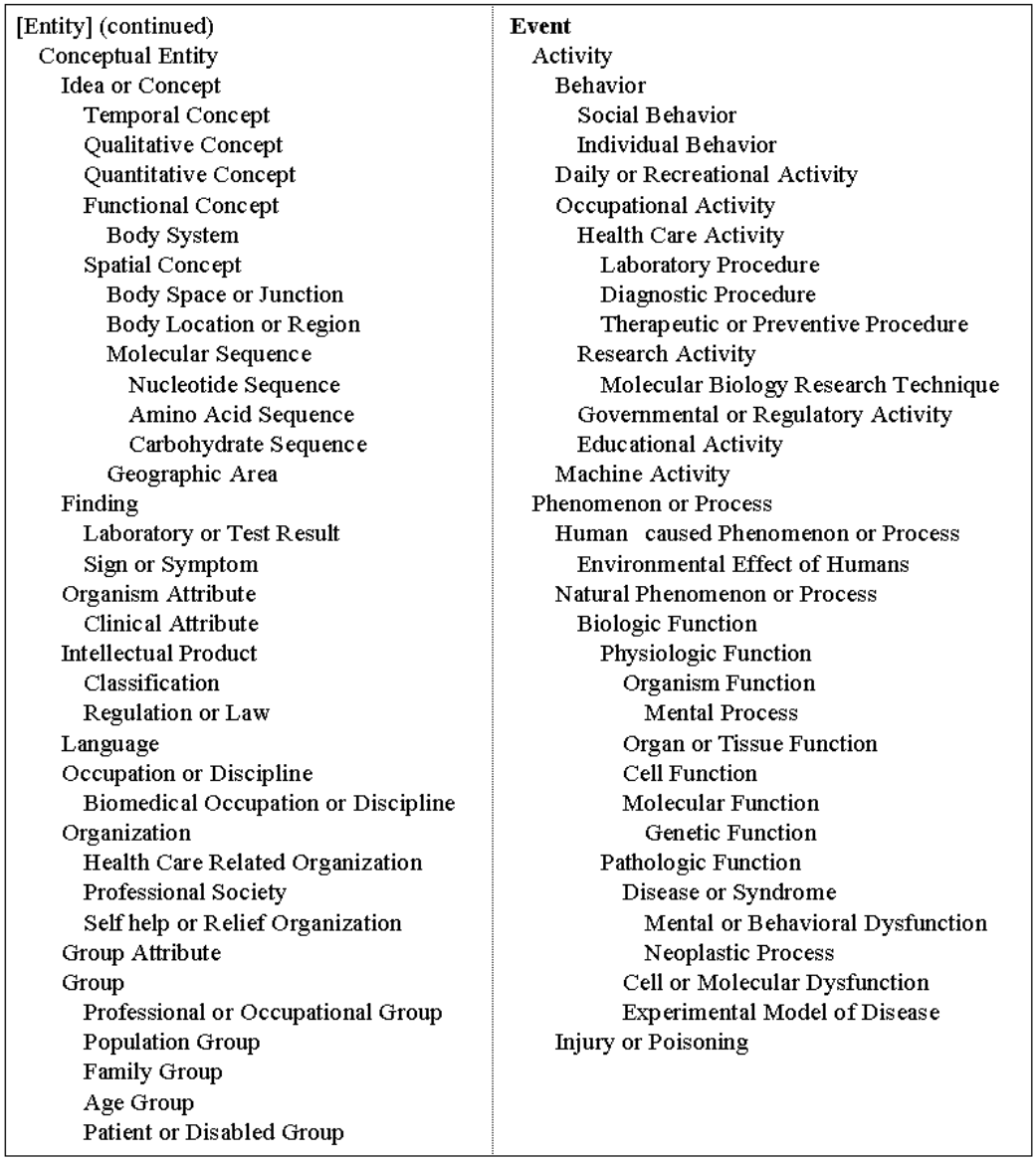
Cada tipo semántico está unido con su hiperónimo por una relación del tipo 'es un(a)' (*is a*), e. g. "Human" es un nodo en la jerarquía de "Entity". La jerarquía que atraviesa las relaciones "is a" desde "Human" hasta "Entity" permite las siguientes relaciones: un "human" es un "mammal", el cual es un "vertebrate"; un "vertebrate" es un "animal", el cual es un "organism"; un "organism" es un "physical object", el cual es una "entity".

A continuación se presentan la figura 6 con la lista de los 134 tipos semánticos de UMLS (2004AB)[39] ordenados de modo jerárquico:

[39] Lista tomada de la versión 2004 AB pero que no ha cambiado en las nuevas versiones de UMLS. Consultada en http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

Entity
- Physical Object
  - Organism
    - Plant
      - Alga
    - Fungus
    - Virus
    - Rickettsia or Chlamydia
    - Bacterium
    - Archaeon
    - Animal
      - Invertebrate
      - Vertebrate
        - Amphibian
        - Bird
        - Fish
        - Reptile
        - Mammal
          - Human
  - Anatomical Structure
    - Embryonic Structure
    - Anatomical Abnormality
      - Congenital Abnormality
      - Acquired Abnormality
    - Fully Formed Anatomical Structure
      - Body Part, Organ, or Organ Component
      - Tissue
      - Cell
      - Cell Component
      - Gene or Genome
  - Manufactured Object
    - Medical Device
    - Research Device
    - Clinical Drug

[Entity] (continued)
- [Physical Object] (continued)
  - Substance
    - Chemical
      - Chemical Viewed Functionally
        - Pharmacologic Substance
          - Antibiotic
        - Biomedical or Dental Material
        - Biologically Active Substance
          - Neuroreactive Substance or Biogenic Amine
          - Hormone
          - Enzyme
          - Vitamin
          - Immunologic Factor
          - Receptor
        - Indicator, Reagent, or Diagnostic Aid
        - Hazardous or Poisonous Substance
      - Chemical Viewed Structurally
        - Organic Chemical
          - Nucleic Acid, Nucleoside, or Nucleotide
          - Organophosphorus Compound
          - Amino Acid, Peptide, or Protein
          - Carbohydrate
          - Lipid
            - Steroid
            - Eicosanoid
        - Inorganic Chemical
        - Element, Ion, or Isotope
      - Body Substance
      - Food



[Entity] (continued)	**Event**
Conceptual Entity	Activity
Idea or Concept	Behavior
Temporal Concept	Social Behavior
Qualitative Concept	Individual Behavior
Quantitative Concept	Daily or Recreational Activity
Functional Concept	Occupational Activity
Body System	Health Care Activity
Spatial Concept	Laboratory Procedure
Body Space or Junction	Diagnostic Procedure
Body Location or Region	Therapeutic or Preventive Procedure
Molecular Sequence	Research Activity
Nucleotide Sequence	Molecular Biology Research Technique
Amino Acid Sequence	Governmental or Regulatory Activity
Carbohydrate Sequence	Educational Activity
Geographic Area	Machine Activity
Finding	Phenomenon or Process
Laboratory or Test Result	Human caused Phenomenon or Process
Sign or Symptom	Environmental Effect of Humans
Organism Attribute	Natural Phenomenon or Process
Clinical Attribute	Biologic Function
Intellectual Product	Physiologic Function
Classification	Organism Function
Regulation or Law	Mental Process
Language	Organ or Tissue Function
Occupation or Discipline	Cell Function
Biomedical Occupation or Discipline	Molecular Function
Organization	Genetic Function
Health Care Related Organization	Pathologic Function
Professional Society	Disease or Syndrome
Self help or Relief Organization	Mental or Behavioral Dysfunction
Group Attribute	Neoplastic Process
Group	Cell or Molecular Dysfunction
Professional or Occupational Group	Experimental Model of Disease
Population Group	Injury or Poisoning
Family Group	
Age Group	
Patient or Disabled Group	

Figura 6: Lista de los tipos semánticos de UMLS.

En la figura 7, se presentan todos los tipos semánticos agrupados en 15 grupos semánticos. En cada grupo semántico, se presenta su codificación en la segunda columna, el número de tipos semánticos y el nombre de cada uno de ello en la última columna. La codificación de la segunda columna de los grupos semánticos se empleará para obtener los patrones semánticos en §6.4.5 y §6.4.7.

Semantic Groups			Semantic Types
Activities & Behaviors	ACTI	9	■ Activity ■ Behavior ■ Daily or Recreational Activity ■ Event ■ Governmental or Regulatory Activity ■ Individual Behavior ■ Machine Activity ■ Occupational Activity ■ Social Behavior
Anatomy	ANAT	11	■ Anatomical Structure ■ Body Location or Region ■ Body Part, Organ, or Organ Component ■ Body Space or Junction ■ Body Substance ■ Body System ■ Cell ■ Cell Component ■ Embryonic Structure ■ Fully Formed Anatomical Structure ■ Tissue
Chemicals & Drugs	CHEM	26	■ Amino Acid, Peptide, or Protein ■ Antibiotic ■ Biologically Active Substance ■ Biomedical or Dental Material ■ Carbohydrate ■ Chemical ■ Chemical Viewed Functionally ■ Chemical Viewed Structurally ■ Clinical Drug ■ Eicosanoid ■ Element, Ion, or Isotope ■ Enzyme ■ Hazardous or Poisonous Substance ■ Hormone ■ Immunologic Factor ■ Indicator, Reagent, or Diagnostic Aid ■ Inorganic Chemical ■ Lipid ■ Neuroreactive Substance or Biogenic Amine ■ Nucleic Acid, Nucleoside, or Nucleotide ■ Organic Chemical ■ Organophosphorus Compound ■ Pharmacologic Substance ■ Receptor ■ Steroid ■ Vitamin
Concepts & Ideas	CONC	12	■ Classification ■ Conceptual Entity ■ Functional Concept ■ Group Attribute ■ Idea or Concept ■ Intellectual Product ■ Language ■ Qualitative Concept ■ Quantitative Concept ■ Regulation or Law ■ Spatial Concept ■ Temporal Concept
Devices	DEVI	2	■ Medical Device ■ Research Device
Disorders	DISO	12	■ Acquired Abnormality ■ Anatomical Abnormality ■ Cell or Molecular Dysfunction ■ Congenital Abnormality ■ Disease or Syndrome ■ Experimental Model of Disease ■ Finding ■ Injury or Poisoning ■ Mental or Behavioral Dysfunction ■ Neoplastic Process ■ Pathologic Function ■ Sign or Symptom
Genes & Molecular Sequences	GENE	5	■ Amino Acid Sequence ■ Carbohydrate Sequence ■ Gene or Genome ■ Molecular Sequence ■ Nucleotide Sequence
Geographic Areas	GEOG	1	■ Geographic Area
Living Beings	LIVB	23	■ Age Group ■ Alga ■ Amphibian ■ Animal ■ Archaeon ■ Bacterium ■ Bird ■ Family Group ■ Fish ■ Fungus ■ Group ■ Human ■ Invertebrate ■ Mammal ■ Organism ■ Patient or Disabled Group ■ Plant ■ Population Group ■ Professional or Occupational Group ■ Reptile ■ Rickettsia or Chlamydia ■ Vertebrate ■ Virus
Objects	OBJC	5	■ Entity ■ Food ■ Manufactured Object ■ Physical Object ■ Substance
Occupations	OCCU	2	■ Biomedical Occupation or Discipline ■ Occupation or Discipline
Organizations	ORGA	4	■ Health Care Related Organization ■ Organization ■ Professional Society ■ Self-help or Relief Organization
Phenomena	PHEN	6	■ Biologic Function ■ Environmental Effect of Humans ■ Human-caused Phenomenon or Process ■ Laboratory or Test Result ■ Natural Phenomenon or Process ■ Phenomenon or Process
Physiology	PHYS	9	■ Cell Function ■ Clinical Attribute ■ Genetic Function ■ Mental Process ■ Molecular Function ■ Organ or Tissue Function ■ Organism Attribute ■ Organism Function ■ Physiologic Function
Procedures	PROC	7	■ Diagnostic Procedure ■ Educational Activity ■ Health Care Activity ■ Laboratory Procedure ■ Molecular Biology Research Technique ■ Research Activity ■ Therapeutic or Preventive Procedure

Figura 7: Lista de los grupos semánticos y los tipos semánticos de UMLS[40].

A continuación se presenta la figura 8 con la lista de las 54 relaciones semánticas de UMLS (2004AB)[41]

[40] Tabla tomada de: Bodenreider, Olivier; McCray, Alexa (2003: 416) *Exploring Semantic Groups through Visual Approaches*. En: *Journal of Biomedical Informatics* 36.

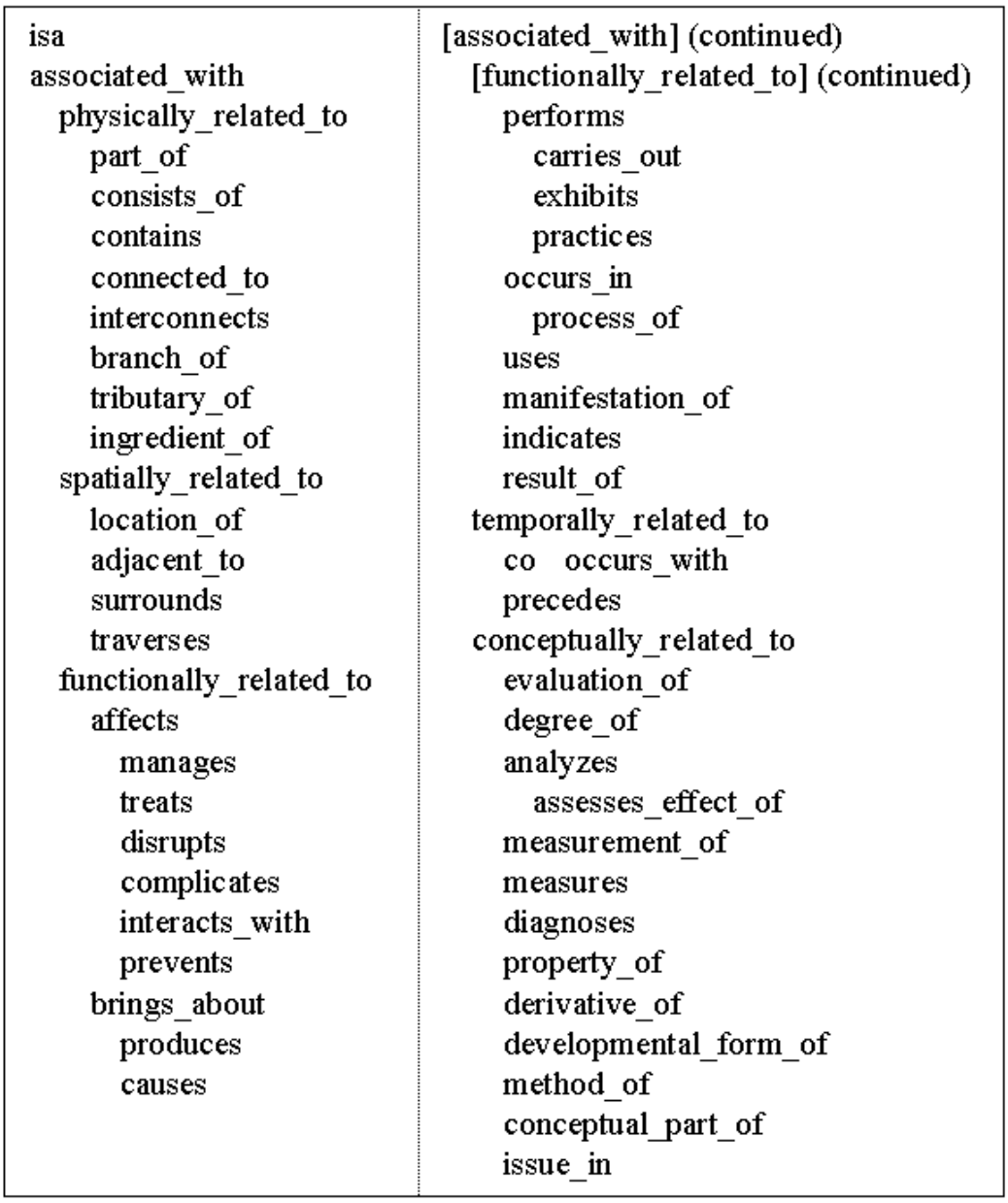


Figura 8: Lista de las relaciones semánticas de UMLS.

[41] Lista tomada de la versión 2004 AB pero que no ha cambiado en las nuevas versiones de UMLS. Consultada en http://www.nlm.nih.gov/research/umls/META3_current_relations.html.

3.2.2 Diccionarios en CD-ROM

Para cualquier consulta o duda, se han empleado los siguientes diccionarios.

Diccionario de lengua española (DRAE)
Gran Diccionario de la Lengua Española (GDLE)
Diccion@rios Espasa
Diccionario Vox
Collins COBUILD
Collins bilingüe inglés-español
Stedman's Medical Dictionary 3.0
Random House Webster's Unabridged

El tipo de consulta más frecuente fue la verificación del significado de una palabra para poder seleccionar la etiqueta adecuada en WordNet, EuroWordNet o UMLS para la comprobación de la forma expandida de una sigla, entre otros.

3.3 Metodología de análisis general

A efectos de claridad, en cada capítulo se describirá la metodología llevada a cabo. Sin embargo, aquí se describirá a grandes rasgos la metodología empleada en toda la tesis como puede verse en el esquema de la figura 9.

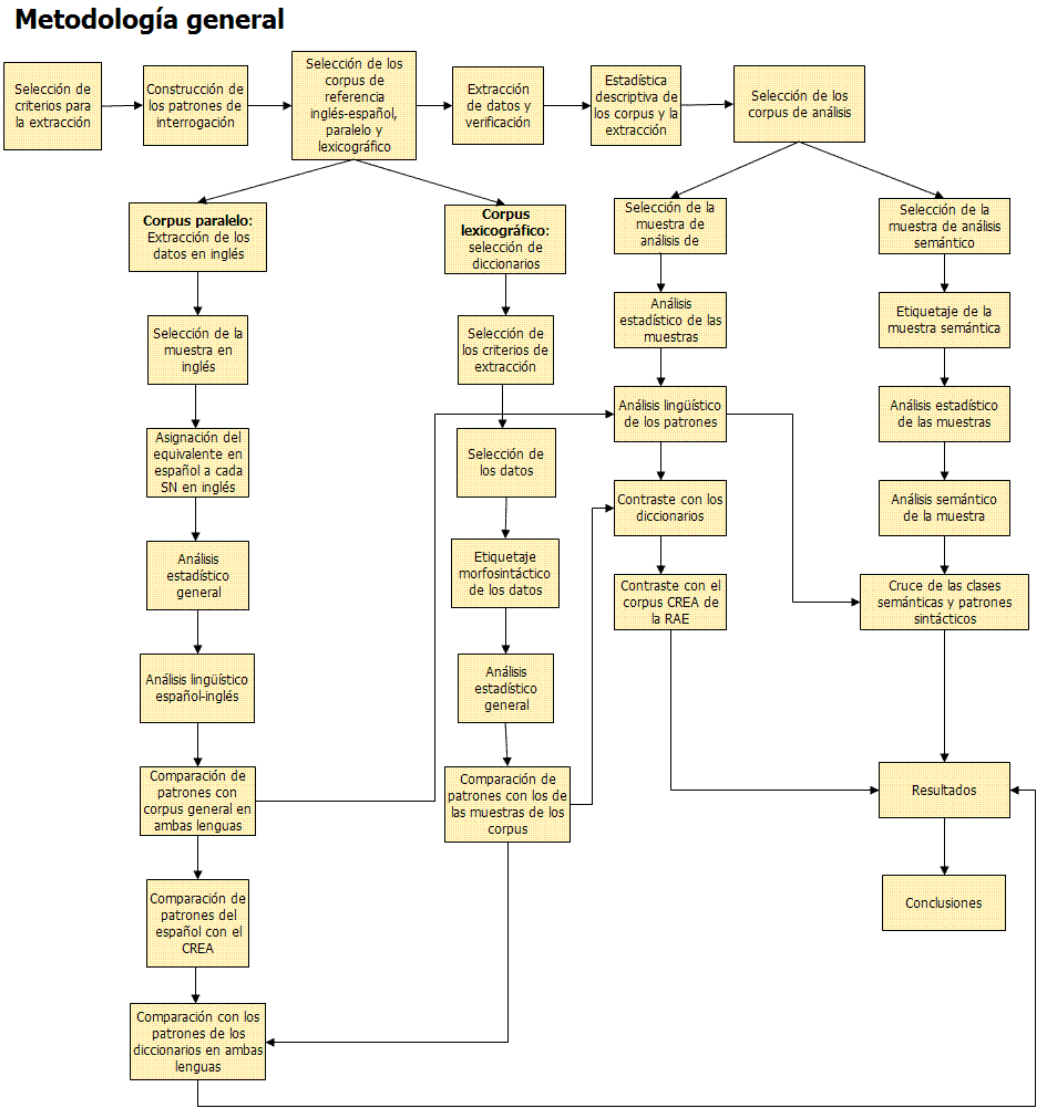


Figura 9: Esquema general de la metodología.

En la primera línea del esquema pueden observarse los pasos principales de la metodología. En primer lugar, el estudio piloto, que además de permitir medir el alcance de un estudio posterior, permitió establecer un conjunto de patrones para posteriormente construir aquellos que permitieran hacer la extracción final de los datos. Posteriormente, se realizó la selección de los corpus de referencia en inglés y en español. Puesto que el procesamiento de corpus paralelo y el corpus lexicográfico se diferencia un poco del procesamiento de los corpus de referencia, se subdivide en el diagrama para poder reflejar esas diferencias. Luego, se realizó la extracción de los datos en los diferentes corpus y fuentes y su estadística descriptiva correspondiente.

Finalmente, se seleccionaron las muestras de los corpus para el análisis formal, semántico y los análisis contrastivos del corpus paralelo y la comparación de los datos contra corpus lexicográfico y el corpus CREA de la RAE.

3.4 Problemas de etiquetaje

Durante el proceso de extracción de los datos del corpus y el proceso etiquetaje semántico, se observaron diversos problemas de etiquetaje.

En primer lugar, mencionaremos los casos sintácticos y, en segundo lugar, mencionaremos los casos semánticos.

En primer lugar, se presentaron categorías léxicas erróneas debido a problemas de desambiguación, principalmente en los participios de pasado y presente en inglés (*-ed, -ing*). Por ejemplo, se presentaron muchos casos en que el sistema desambiguó mal un sustantivo deverbal como *test* y lo etiquetó como verbo. De igual modo, etiquetó participios de presente en vez de sustantivos y participios de pasado como verbos en pasado. De este modo, se refinaron todos los patrones con las extracciones sucesivas de los sintagmas nominales hasta poder obtener los resultados esperados en la fase de extracción.

Para poder recuperar patrones que contenían este tipo de categorías se procedió entonces a crear patrones falsos. En el ejemplo 1 y 2 se construye un patrón en inglés y español, respectivamente con un verbo como núcleo y en el ejemplo 3 se confecciona 1 patrón con adjetivo como sustantivo nuclear.

1.	H6 + X + V.* (PP X V)	altered malf-phoa fragment
2.	JQ.* + V.* + JQ.* (Adj V Adj)	buen estado general
3.	JQ.* + VC.* + P + N5.* (Adj V Prep N)	pacientes diagnosticados de depresión

En español, se dieron casos como en el ejemplo 3, en el cual el sustantivo *paciente* estaba etiquetado como adjetivo. Igualmente, el sistema de etiquetaje contiene un lexicón general y, debido a que se trabajó con un corpus especializado, se presentaron muchos casos de términos y palabras que no fueron reconocidos por el etiquetador y, por tanto, fueron marcados con la categoría X en inglés o W en español para indicar que dichas palabras no están en el lexicón del sistema. Por tanto, se tuvo que crear también patrones falsos para poder recuperar sintagmas de este tipo, como se describe en los ejemplos 4, 5 y 6.

4.	JA + JA + X	(Adj Adj X)	bilateral central epileptiform
5.	X + X + NN.*	(X X N)	laser-desorption time-of-flight mass
6.	X + X + NN.*	(X X N)	calcium-modulating cyclophilin ligand

Esto corrobora la afirmación de Maniez (2001: 56) en cuanto a que el propósito de los etiquetadores está dirigido más a la lengua general y no a ámbitos especializados.

En segundo lugar, se presentaron problemas en el etiquetaje semántico de determinados lemas o formas. En otros casos, existe el *synset* en WordNet pero pertenece a una temática diferente y no concuerda ese sentido el área con una de las áreas nuestras como sucede con el caso de “tasas” en “altas tasas de mutación”, en el cual WordNet le asignas las áreas *tax*, *money* y *economy* pero no estadística que sería más adecuada en este caso.

Existen otros casos en los cuales el significado del lema no corresponde exactamente al significado dentro del sintagma. Por ejemplo, el lema “transmisión” en el sintagma “transmisión autosómica recesiva”, se refiere en WordNet al “acto de enviar un mensaje” y no al acto de pasar información.

Existen casos como en *horizontal* en los cuales WordNet le asigna como adjetivo la clase *noun.attribute* pero en el significado de sustantivo le asigna la

clase *noun.relation* que es más adecuada en el caso del sintagma *horizontal gene transfer*.

Para buscar el *synset* de muchos adjetivos o adverbios hubo que ampliar la búsqueda a *Adj.+Derivational related forms* o *Synonyms related nouns*, lo que dificultó y multiplicó el tiempo de búsqueda. Por ejemplo, si se busca el adverbio *anatomically*, WordNet 2.1 no nos proporciona la información del *synset*, como se presenta en la figura 10.

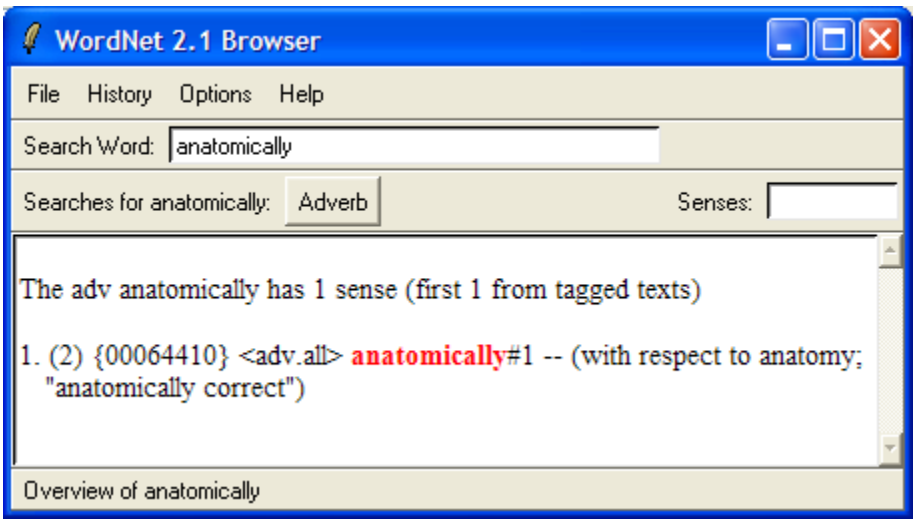


Figura 10: Forma de mostrar la información de los adverbios en WordNet.

Si se extiende la búsqueda por ***Synonyms/Stem Adjectives*** con el botón ***Adverb***, no se puede recuperar el *synset* como se ve en la figura 11. WordNet sólo nos muestra que proviene de un adjetivo relacional (*adj.pert*).

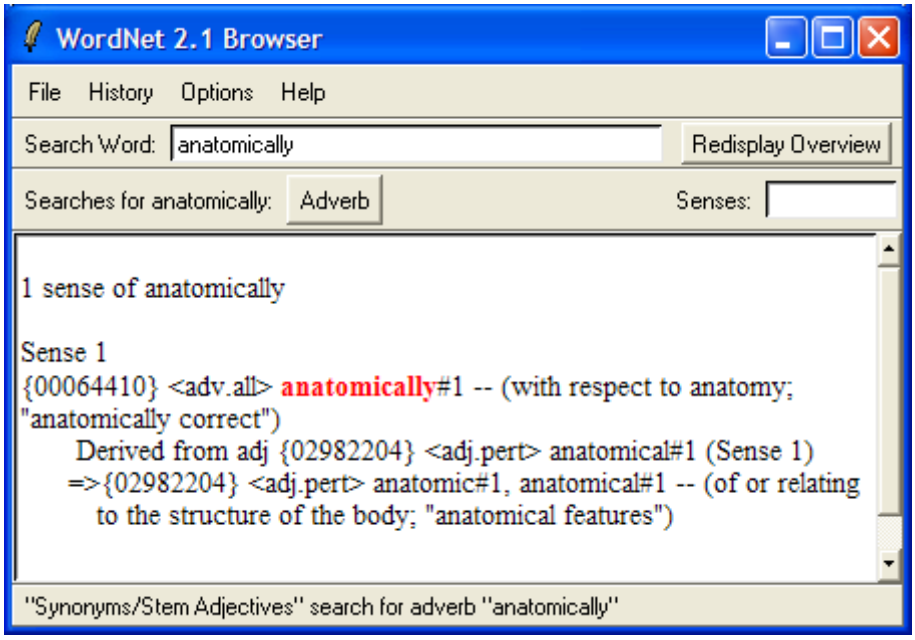


Figura 11: Forma de mostrar la información relacionada de los adverbios en WordNet.

Por tanto, para poder recuperar el *synset* del adverbio *anatomically* es necesario buscar por el adjetivo *anatomical* y la combinación de ***Derivational related forms*** en el botón ***Adjective,*** como se observa en la figura 12.

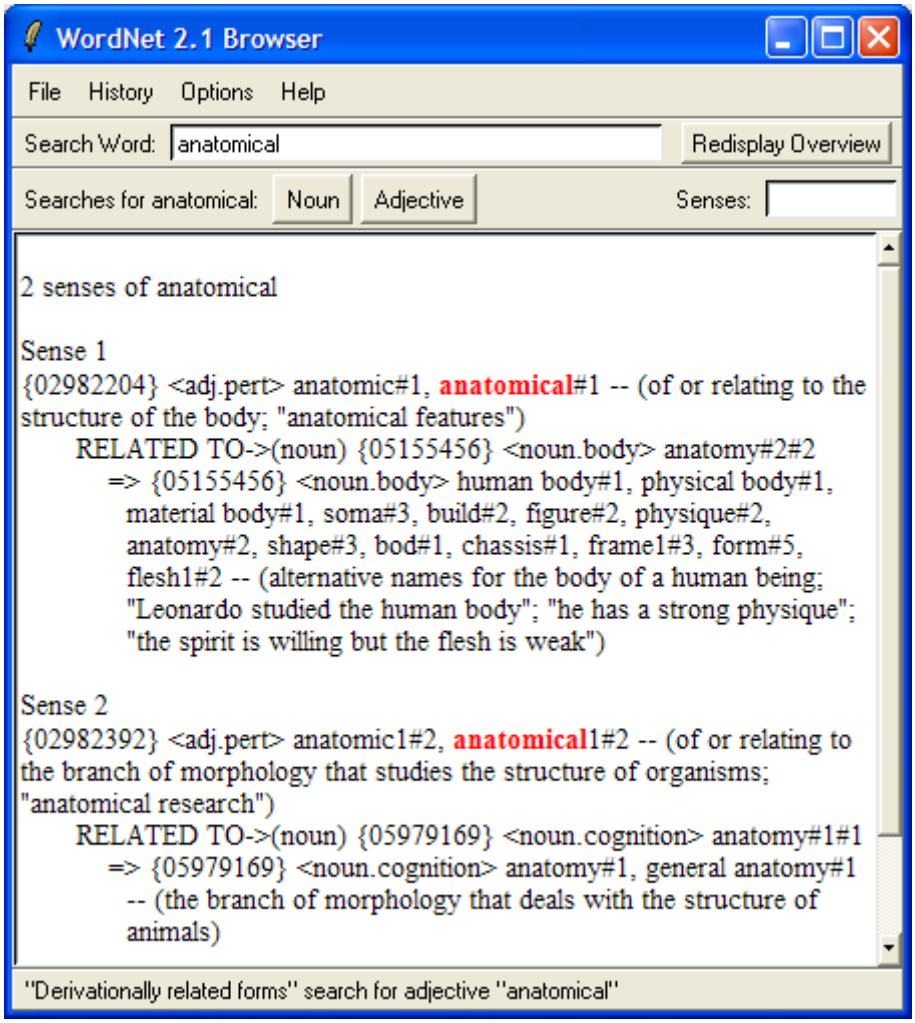


Figura 12: Forma de recuperar el *synset* de los adverbios mediante el adjetivo en WordNet.

3.5 Extracción de las unidades y tratamiento de los datos

A partir de los resultados obtenidos en Quiroz (2003), y en especial en Quiroz *et al* (2004), además de la observación de trabajos que incluyen unidades similares, se decidió extraer todas las unidades que, en general, cumplieran con el criterio de tener al menos dos premodificadores de categoría abierta, es decir, sustantivos, adjetivos, participios (*-ing*, *-ed*) y adverbios. Las otras categorías no se tuvieron en cuenta para el análisis cuantitativo, pero sí

para otros tipos de análisis. De igual modo, en español no se contaron algunas categorías frecuentes en los sintagmas, como los determinantes que están a la izquierda del núcleo ni las preposiciones.

Por otro lado, se extrajeron todos los sintagmas nominales que simplemente cumplían el requisito de extensión sin importar su grado de lexicalización (para mayor referencia véase la cita de L'Homme en §2.2. Por tanto, para el análisis cuantitativo se tuvieron en cuenta sintagmas endógenos y exógenos, compuestos (*compounds*) del tipo N+N+... N, etc.

La extracción se realizó con base en los patrones de superficie de las reglas de entrenamiento obtenidos en los trabajos del DEA mediante la herramienta Repoker[42]. Para completar los patrones del estudio piloto, se revisó la literatura en inglés y en las lenguas romances de modo que no se dejaran patrones potenciales por fuera. Se incluyeron patrones en inglés estudiados o mencionados tangencialmente por Montero (1995), Biber *et al* (1999), Collet (2003), entre otros.

De igual modo, se revisó la literatura en español y en otras lenguas romances (catalán, francés y portugués) para tomar patrones ya estudiados o mencionados por Montero (1995), Estopà (1999), Cartagena (1999), Café (1999), Collet (2003), Cardero (2004) y Vilvaldi (2004).

Al final, se coleccionaron y crearon 99 patrones en español y 50 patrones en inglés[43] con los cuales se procedió a realizar la extracción en los dos corpus de referencia del *Corpus Tècnic* del IULA y posteriormente, en el corpus paralelo. Con un *script* de Perl, a cada patrón se le asignó el número de tokens y el patrón con la codificación del *Corpus Tècnic* del IULA. En esta extracción, se obtuvo un total de 21.521 sintagmas en inglés y 38.424 en español sin repetir de

[42] Herramienta de extracción creada en Perl por el Dr. de Yzaguirre del IULA en el año 2004.

[43] El listado completo de patrones aparece en los anexos 1 y 2.

los corpus de referencia y 1.694 SN en inglés del corpus paralelo como se presenta en la tabla 5.

	Inglés	Español
Corpus de referencia	21.521 SN	38.424 SN
Corpus paralelo	1.694 SN	x[44]

Tabla 5: Primera extracción de sintagmas realizada en los corpus de referencia.

Se presentaron diversos problemas durante el proceso de extracción de los sintagmas. El primer caso, sucede cuando un patrón subsume a otro y, por ende, un sintagma se extrae dos o más veces erróneamente como sucede con el sintagma *new human mitochondrial atp-binding cassette membrane protein* del ejemplo 7.

7. new human mitochondrial atp-binding cassette membrane protein
Adj Adj Adj N N N N

new human mitochondrial atp-binding cassette membrane
Adj Adj Adj N N N

new human mitochondrial atp-binding cassette
Adj Adj Adj N N

new human mitochondrial atp-binding
Adj Adj Adj N N

ATP-binding cassette membrane
N N N N

cassette membrane protein
N N N

---

[44] La cantidad de sintagmas del corpus paralelo no aparece ya que este corpus se empleó para colocar los sintagmas equivalentes a la muestra de corpus paralelo en inglés.

El segundo caso ocurre cuando un sintagma con dos premodificadores unidos por un guión en el corpus será de dos tokens. No se podrá extraer con las reglas actuales pues no se han creado patrones de dos tokens, como sucede con los ejemplos 8 y 9.

8. AIDS-related death (X N)
9. placebo-controlled trial (X N)

Como se comentó antes, no se extrajeron los sintagmas coordinados con y, o coma, dado que no era posible desambiguar manualmente todos los casos como en los ejemplos 10, 11 y 12 y su análisis formal presenta otra serie de opciones que no se tenían previstas.

10. a specific and potent inhibitor
11. this randomised, double-blind, placebo-controlled, multicentre trial
12. routine haematological and biochemical laboratory studies

Se presentaron casos falsos de etiquetaje como en los ejemplos 13, 14 y 15 los cuales están etiquetados como X, es decir, como unidades que no están incluidas en el lexicón del sistema. Lo sorprendente de estos casos es que simplemente son errores de etiquetaje pues en otros ejemplos se puede recuperar la categoría gramatical correcta.

13. which viral dna (X Adj N)
14. also competitive inhibitor (X Adj N)
15. have high affinity (X Adj N)

Para poder solucionar este ruido, se procedió a crear restricciones en todos los patrones de búsqueda con el ruido localizado en las extracciones previas como se ve en el ejemplo 16 para recuperar sintagmas como *highly activated myofibroblastic cells* con el patrón “D6 H6 X NN.*” (Adv PP N N) y volver a realizar la extracción.

16. [pos="**D6**" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"]
[pos="**H6**"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"]
[pos="**X**"&lemma!="that|which|who"]
[pos="**NN**.*"]

En la restricción se le ha pedido que recupere adverbios y, en especial con las unidades terminadas en *–ly* y unidades como *in vivo*, *in vitro*, etc. Además, se le ha pedido que recupere participios de presente y pasado y que no incluya palabras tales como *containing, including, having*, etc. De este modo, se asegura que no se recuperaran sintagmas con formas que son verbales y no se excluyeron participios de presente que si son parte de sintagmas como en *corresponding cloned cdna*. Con la categoría gramatical X, se recuperaron unidades que no están en el lexicón del sistema, en especial, unidades muy especializadas y además se ha restringido a las unidades que darían ruido como *that*, *which*, *who*, entre otros casos como en los ejemplos 13-15 antes descritos.

De igual modo, se presentaron problemas con unidades marcadas como XXX como en los ejemplos 17, 18 y 19. Estas son unidades (siglas, símbolos, números, etc.) que se han eliminado en el preproceso del texto por diversas razones que no competen a esta tesis.

17.

mitochondrial	mitochondrial\JA
xxx	xxx\MO6
protein	protein\NN6S

18.

human	human\JA
ovarian	ovarian\JA
xxx	xxx\MO6
cells	cell\NN6P

19.	genomic	genomic\JA
	xxx	xxx\MO6
	DNA	dna\NN6S
	clone	clone\NN6S

Como consecuencia, se perdieron muchos sintagmas con patrones debido al etiquetaje y que potencialmente podían tener una de las unidades antes descritas. En el caso de los sintagmas más extensos esta dificultad era aún más evidente y probable debido a que este tipo de discurso suele tener muchas siglas, números y nomenclaturas.

3.6 Aspectos estadísticos

Para la clasificación de los datos y el análisis estadístico de cada uno de los corpus, se empleó la siguiente estrategia.

En primer lugar, se clasificaron y contaron los sintagmas de cada lengua de menor a mayor extensión con su respectivo patrón. Luego, se calcularon los porcentajes por número de palabras de categoría léxica abierta (3, 4, 5, 6, 7, y 8 palabras) y, luego, contra el número total de palabras (vocabulario). Así, se quería verificar si la tendencia observada en Quiroz (2004, 2005, 2006) continuaba, es decir, se quería corroborar si el porcentaje de sintagmas extensos es mayor que en los otros estudios mencionados en §2. Con ello, se reforzará la hipótesis desde un punto de vista cuantitativo, de que este fenómeno es "natural" en este tipo de texto y que su presencia es un rasgo que lo caracteriza.

En segundo lugar, se extrajeron los patrones en un fichero para ser analizados en el programa Statgraphics, el cual permite obtener las regularidades de los patrones de superficie y de las diferentes categorías. Se calcularon los porcentajes de las frecuencias de los patrones en general y se calcularon los porcentajes de las frecuencias de los patrones por predominio de

categoría y por exclusión de categoría, es decir, se clasificaron los patrones que tienen N, A, Adv, PP, PPi en la premodificación y también, de forma excluyente, patrones que tienen N en la premodificación pero no A y patrones que tienen A en la premodificación pero no N. De este modo, se definirán los patrones de superficie más frecuentes, la categoría gramatical que predomina en la premodificación y el comportamiento del resto de categorías en los patrones.

Para el corpus paralelo, se siguió el mismo procedimiento, pero siempre ordenando ambas lenguas al mismo tiempo para poder observar luego las regularidades en las traducciones respecto del inglés.

Para poder mantener un control adecuado de los datos, se emplearon diversas hojas en el programa Excel 2003. Las tablas o datos que se filtraban en Excel se tabulaban en el programa Statgraphics.

Para el análisis morfológico se prepararon las diferentes listas de acuerdo con el sufijo.

Las dependencias sintácticas se marcaron con números y posteriormente se cruzaron con los patrones superficiales para obtener las diferentes tablas presentadas en §6 y §7. Una vez obtenidas las tablas, los números de las dependencias se convirtieron a una estructura de corchetes tipo [A [B C]].

En la tabla 6 se muestra un ejemplo de un sintagma con sus etiquetas morfosintácticas y semánticas (mapeadas posteriormente). Una vez etiquetados semánticamente cada uno de los tokens de las muestras, se introdujeron en Statgraphics para obtener la tabulación de las diferentes clases en WordNet, EuroWordNet o UMLS.

ID	Tokens del SN	N.º del elemento	Cat. léxica	WordNet 2.1	UMLS Semantic type
26	anatomically	3	D6	adv.all//adj.pert/noun.body	Functional Concept
26	modern	2	JA	adj.all/noun.attribute	not found
26	humans	1	NN	noun.animal	Human

Tabla 6: Datos en Excel con la información léxica y semántica.

Para poder obtener los patrones superficiales y semánticos, se han paralelizado hasta obtener las secuencias presentadas en la tabla 7.

ID	Sintagma nominal	Patrón superficial	Patrón WordNet 2.1	Patrón UMLS
26	anatomically modern humans	D6 JA NN	noun.body noun.attribute noun.animal	Functional Concept not found Human

Tabla 7: Datos en Excel con los patrones superficiales y semánticos.

Una vez se han obtenido los patrones, se introdujeron en Statgraphics para tabular todos los patrones superficiales y semánticos en los tres programas empleados.

En el caso de UMLS, se mapearon todos los tipos semánticos a los nueve grupos semánticos con el fin de obtener más regularidades siguiendo el procedimiento anterior. Por ejemplo, el sintagma *anatomically modern humans* tiene los tipos *Functional Concept not found Human* que se han mapeado a los grupos semánticos CONC NotF LIVB.

3.7 Selección de las muestras para los análisis

A partir de los 50 patrones en inglés y los 99 patrones en español, se extrajeron 21.521 sintagmas en inglés y 38.431 sintagmas en español.

Con base en la extracción realizada, se determinó una muestra de 1.060 SN para el inglés y 1.087 para el español con un error del 3% para realizar los análisis que se presentan a partir de §4. Este tamaño muestral se distribuyó proporcionalmente de acuerdo con la frecuencia de aparición de cada patrón. Para seleccionar la muestra, se descartaron todos los patrones con una frecuencia menor a 5 ocurrencias lo que redujo considerablemente el número de

patrones elegibles a 33 en inglés y 60 en español[45]. Puesto que se debía seleccionar unidades completas se redondeó cada cifra a un número inferior o superior. Así la muestra sintáctica final quedó distribuida de la siguiente manera: 1.055 sintagmas para el inglés y 1.096 sintagmas para el español como puede verse en la tabla 6.

	Extracción total	Error aprox.	Tamaño calculado de la muestra	Tamaño final de la muestra	Patrones de la muestra final
Inglés	21.521	3%	1.060	1.055	33
Español	38.424	3%	1.087	1.096	60

Tabla 6: Resumen de sintagmas extraídos, muestra sintáctica y número de patrones.

Para la selección de la muestra para los análisis de dependencias y semántico, se calculó una muestra aproximada del 20% a partir de la muestra sintáctica con los 10 patrones más frecuentes de la muestra.

Para el inglés, se seleccionaron 232 SN, un 24,37% de los 1.055 de la muestra sintáctica y para el español 200, un 22% de los 1.096 sintagmas de muestra.

Toda la muestra semántica se etiquetó manualmente con WordNet 2.1 y UMLS para el inglés y con EuroWordNet 1.6 para el español.

	Muestra morfosintáctica	Muestra de dependencias y semántica	Porcentaje
Inglés	1.096	232 SN	24,37%
Español	1.055	200 SN	22%

Tabla 7: Muestra de sintagmas extraídos para los análisis de dependencia y semántico.

En cuanto a los textos paralelos, se extrajeron 1.649 sin repetir representados en 157 patrones de superficie. Se seleccionó una muestra de 332

[45] Para ver la lista completa de patrones finales con ejemplos, véase los anexos 3 y 4.

sintagmas con un error aproximado del 5% para asignar a cada sintagma su equivalente en español.

	Extracción total	Patrones	Error aprox.	Tamaño de la muestra calculado	Tamaño final de la muestra
Inglés	1.649	157	5%	320	332
Español				320	332

Tabla 8: Muestra de sintagmas extraídos para el análisis del corpus paralelo.

Todas las muestras se separaron manualmente de la extracción inicial y todos los sintagmas se extrajeron manualmente bajo los siguientes criterios:

- frecuencia de mayor a menor
- carácter terminológico del núcleo y sus modificadores si fuera posible
- corrección del sintagma (completo)

Finalmente, a cada sintagma de la muestra se le asignó su equivalente en español y el patrón superficial correspondiente. En cada capítulo se explicarán los criterios, la metodología y los análisis correspondientes.

4. Análisis formal de los patrones en inglés

4.1 Introducción

La premodificación es una función sintáctica del sintagma nominal y puede tener varios tipos de modificadores: el adjetivo como en el ejemplo 1, los participios de presente y pasado como en los ejemplos 2 y 3, y el mismo sustantivo como en el ejemplo 4.

1. **total human genomic** DNA
2. **inherited** mitochondrial DNA diseases
3. **circulating** monoclonal protein
4. vehicle control **cell** growth

De igual modo, se pueden encontrar otras categorías léxicas e incluso estructuras gramaticales en la premodificación. En el ejemplo 5 se presenta un sintagma adverbial que, en su conjunto, modifica al núcleo. En el ejemplo 6 aparece una forma verbal como premodificador y en el ejemplo 7 un sintagma preposicional.

5. **darkly** stained apical dendrites
6. **need-to-know** basis
7. **after-sales** manager

En cuanto a su distribución, la premodificación tiende ser más común en inglés que la posmodificación (Biber *et al* 1999: 578) y en el registro académico, al menos un 60% de los sintagmas nominales tiene algún premodificador. La premodificación es semánticamente menos explícita que la posmodificación para identificar las relaciones entre los premodificadores y el núcleo, debido a la ausencia de preposiciones que son las que mantienen dichas relaciones (Varantola 1984: 38). Lo anterior se hace más evidente cuando aumenta la extensión del sintagma (Trimble 1985: 133; Vivanco 1994: 755), como en los

sintagmas con premodificación compleja que se observan en esta tesis. Sin embargo, esta una estructura más compacta y densa ya que en poco espacio contiene una gran cantidad de información de forma precisa (Varatola 1984: 43).

Algunos autores han informado tangencialmente la presencia de sintagmas con premodificación compleja, en muchos casos, más como una característica rara del lenguaje (Orwellian 1974, Quirk *et al* 1985, Trimble 1985, Burnett 1992, Huckin 1991, Kirkman 1992, Blake y Bly 1993, Norman 1999, entre otros). Biber *et al* (1999: 597) muestran cuantitativamente su presencia en diferentes registros discursivos.

La cuestión sintáctica actual de la premodificación se centra en el orden preferido y las restricciones sintácticas de la premodificación. Quirk *et al* (1985: 1341 y ss) dividen el análisis del orden de la premodificación en cuatro zonas: *precentral*, *central*, *postcentral* y *prenuclear* (*pre-head*[46]). Además, se subdividen como se explica a continuación en el gráfico 1.

La primera zona, denomina *precentral*, incluye los adjetivos no graduables (*non-gradable adjectives*), entre ellos, los adjetivos intensificadores (*intensifying adjectives*):

- enfatizadores (*emphasizers*): certain, definitive, plain, pure
- amplificadores (*amplifiers*): absolute, entire, extreme, perfect, total
- atenuadores (*downtoners*): feble, slight

En la segunda zona denominada *central* se encuentran los adjetivos de grado (*gradable adjectives*) (*big, powerful, slow, thick*), es decir, los adjetivos prototípicos para aplicar el test de su valoración (uso atributivo en el sintagma, uso predicativo en la oración y modificación por '*very*'). Estos se dividen en los

[46] Se ha dejado en muchos casos el nombre en inglés pues la traducción de algunos de ellos es problemática en español.

adjetivos no derivativos (*nonderived adjectives*) (*big, powerful, slow*) y en los adjetivos derivativos (*derived adjectives*): deverbales (*interesting, interested, hesitant*) y los denominales (*angry, rainy, peaceful*).

Según estos autores, el orden de la segunda zona en caso de coocurrencia corresponde a la estructura '*nonderived+deverbal+denominal*'. Además, los adjetivos de tamaño, longitud y altura anteceden a los adjetivos no derivativos (*nonderived adjectives*). El grupo de adjetivos emotivos, evaluativos o subjetivos (*lovely, nice, wonderful, terrible, horrible*) suelen preceder a los otros adjetivos centrales.

En la tercera zona, la *postcentral*, se pueden encontrar los participios de presente y pasado y los adjetivos de color.

La cuarta zona, la *prenuclear,* contiene los premodificadores que menos se acercan al adjetivo, pero que tienen más carácter nominal. Esta zona puede dividirse en tres subzonas:

- adjetivos con propiedades de nombres propios que denotan nacionalidad, origen y estilo (*American, Gothic*)
- adjetivos con características morfológicas o semánticas relacionadas con sustantivos y que tienen el significado de '*consisting of*' o '*relating to*' (*annual, economic, medical, social, political, rural*)
- sustantivos

Si dos premodificadores de la misma clase coocurren, los adjetivos que denotan lugar/tiempo deben ir antes (*local economic interest, annual linguistic meeting*); éstos normalmente no pueden coordinarse.

Para Quirk *et al* (1985: 1341), el principio general para ordenar los modificadores es la polaridad subjetivo/objetivo: las propiedades inherentes al núcleo, visualmente observables y objetivamente reconocibles o accesibles,

tienden a ir más cerca del núcleo. Si el adjetivo es una cuestión de opinión, que no se puede observar visualmente, tenderá a ir más lejos del núcleo.

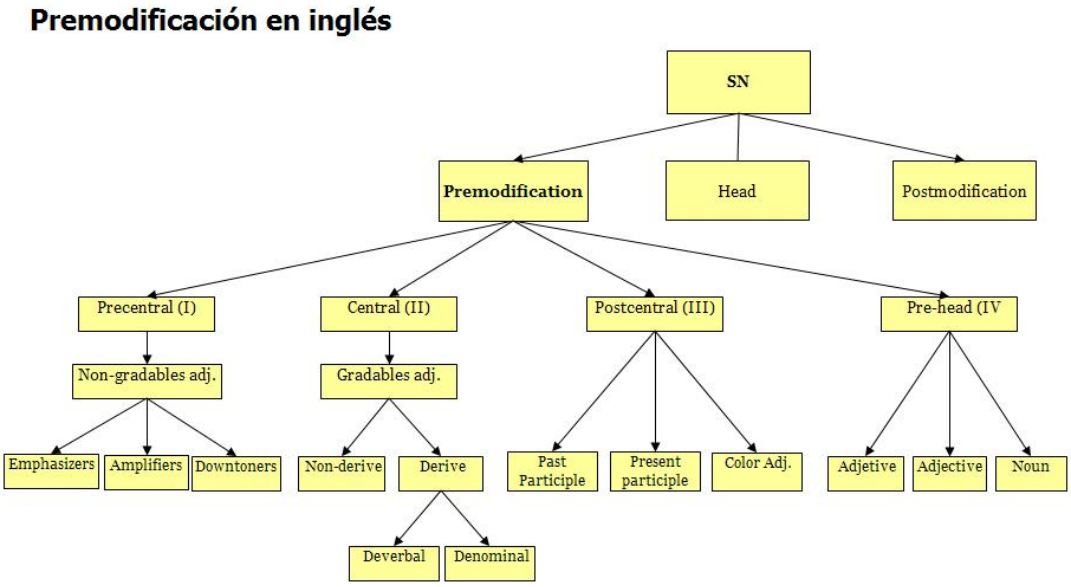


Gráfico 1: Orden de la premodificación en inglés según Quirk *et al* 1985.

Según el gráfico anterior, los modificadores del sintagma del ejemplo 8 ocupan las siguientes zonas:

8.

[subsequent	[placebo-	controlled]	[clinical	[trials]]]]
A	N	PP	A	N
I	III		IV	

Huddleston y Pullum (2002: 452-454) distinguen básicamente sólo dos grandes zonas de análisis en la premodificación.

Pre-head internal modifiers > Head

Huddleston y Pullum (2002: 452-454) distinguen dos grupos dentro de los modificadores internos prenucleares: modificadores internos prenucleares primarios (*early pre-head internal modifiers*) y modificadores internos

prenucleares residuales (*residual pre-head internal modifiers*). En los modificadores internos prenucleares primarios (*early pre-head internal modifiers*) se encuentran los determinantes, superlativos y adjetivos primarios.

Los '*residual pre-head internal modifiers*' se han dividido en las siguientes clases, como se aprecia en el ejemplo 9 tomado de Huddleston y Pullum (2002: 452).

9.

	Evaluative>	General property>	Age>	Colour>	Provenance>	Manufacture>	Type	
an	[attractive	tight-fitting	brand-new	pink	Italian	lycra	women's]	swimsuit

Cada clase de las anteriores puede incluir una subclasificación de modificadores del siguiente tipo, como se ejemplifica en 9:

Evaluativos: modificadores que incluyen la evaluación del hablante y no una propiedad general objetiva (*good, bad*)

- Propiedades generales: tamaño, dimensión, sonido, tacto, gusto, etc.
- Edad
- Color
- Procedencia
- Manufactura: modificadores de composición, modo, etc.
- Tipo

Así pues, el principio de polaridad también está presente en la propuesta de estos autores.

Las preferencias de orden expuestas anteriormente (Quirk *et al* 1985; Huddleston y Pullum 2002), deben estar presentes también en el discurso especializado. Es decir, las restricciones en el orden y el principio de polaridad deberán verse reflejados en los resultados que se presentan en este capítulo.

4.2 Criterios y selección del corpus de análisis en inglés

Como se explicó en el capítulo de la metodología (§3.7), se extrajo una muestra con un 3% de error del corpus de análisis en inglés con 1.055 ocurrencias para el corpus de análisis de los patrones superficiales. Cada sintagma de la muestra se seleccionó manualmente y se distribuyó proporcionalmente de acuerdo con la frecuencia de aparición de cada patrón. Por ejemplo, el patrón más frecuente fue N N N con más de 6.477 ocurrencias en el corpus total y, por tanto, le correspondía una muestra de 317 sintagmas de los 1.055 sintagmas totales. En cambio, el patrón con menos frecuencia es el PP PP N (22) y le correspondía 1 sintagma de muestra.

Asimismo, dentro de cada patrón se seleccionaron, en primer lugar, los sintagmas de mayor a menor frecuencia hasta completar la muestra que correspondía a cada patrón sintáctico. Por ejemplo, en el patrón N N N, el sintagma más frecuente es *polymerase chain reaction* con 37 ocurrencias y el menos frecuente es *tumour necrosis factor* con 2 ocurrencias.

A esta muestra se le realizó una estadística descriptiva en cuanto a la distribución por longitud de los sintagmas en el corpus, categoría léxica predominante, patrones más frecuentes en la muestra, patrones más frecuentes por extensión y patrones más frecuentes por categoría léxica.

Posteriormente, se analizó el corpus lexicográfico de contraste en inglés bajo los mismos parámetros que con la muestra de analisis del corpus. Sin embargo, en cada análisis cuantitativo se comparan los cinco diccionarios entre sí.

Finalmente, se comparan los datos obtenidos en la muestra con los del corpus lexicográfico de contraste y se analizan a la luz de los resultados obtenidos por otros autores en algunos de los análisis llevados a cabo aquí.

Por tanto, se pretende en este capítulo demostrar que los sintagmas con premodificación compleja no son un fenómeno raro de la lengua inglesa. Este hecho se ve reflejado tanto en el corpus de análisis como en el corpus lexicográfico de contraste. Por un lado, las tendencias de los patrones no sólo son cuantitativas sino cualitativas; y, por el otro, la explicación lingüística en el marco de la lengua general, así como su motivación pragmática están mediadas por la interacción de los interlocutores del discurso.

4.3 Resultados

De acuerdo con los criterios descritos anteriormente, se han tabulado los datos a partir del programa de estadística Statgraphics Plus 5.1 para obtener los resultados.

4.3.1 Longitud y frecuencia de los SN en inglés

En el corpus de análisis, se extrajeron unidades desde 3 tokens (uno como núcleo) hasta 8 tokens, como se ve en los ejemplos 10 y 11.

10. T4 polynucleotide kinase (N N N)
11. human acute lymphoblastic leukemia ccrf-cem cdna library (Adj Adj Adj N N N N)

En estudios previos (Quiroz 2005), se encontraron unidades aún más extensas. Estas unidades se extrajeron manualmente lo que ha representado problemas de etiquetaje y extracción, tal y como se ha comentado en §3.7, en

relación con las unidades más extensas. De igual modo, Cortés (2004) extrajo unidades de hasta 14 tokens (*GEM'S auto volume control pto shaft speed related system*).

Como puede verse en la tabla 1, los patrones de 3 tokens (dos en la premodificación) son los más frecuentes en la muestra (corpus de análisis) con un 86,16% del total (909 ocurrencias). Por el contrario, los sintagmas de 4 y 5 tokens tan sólo representan un 13,84% del total de sintagmas (12,8% y 1,04%, respectivamente).

N.º tokens	Frecuencia	Porcentaje
3	909	86,16
4	135	12,8
5	11	1,04
Total	1055	100

Tabla 1: Frecuencia por número de tokens del corpus de análisis del inglés.

En el corpus general, también se extrajeron patrones iguales o mayores a seis tokens, pero no se han tenido en cuenta por su baja frecuencia, ya que el criterio de inclusión fue de más de cinco ocurrencias en la muestra.

Los estudios descritos en §2 presentan la misma tendencia que el corpus de esta tesis, es decir, a menor extensión del patrón, mayor frecuencia de aparición. No es casualidad que la muestra sólo contenga patrones de 3, 4 y 5 tokens ya que estas estructuras son las que pueden revertir más en estabilización y posible lexicalización del sintagma, como lo plantea Cartagena (1998).

4.3.2 Categoría léxica predominante en la premodificación

En principio, la categoría léxica por excelencia para modificar el sustantivo debe ser el adjetivo. Así se deriva de su función atributiva no sólo en la oración sino en el propio sintagma.

Como se explicó en §2.3, Biber *et al* (1999: 589) proponen que el adjetivo es la categoría léxica más frecuente en el "discurso académico". Sin embargo, en un estudio previo (Quiroz *et al* 2004), se observó que aparentemente el sustantivo era la categoría léxica más frecuente en la premodificación y no el adjetivo (42% vs. 17%, respectivamente). Como puede verse en la tabla 2, esta tendencia continúa en este corpus de análisis, aunque la diferencia es menor ya que hay más adjetivos en este corpus (45,95% vs. 32,43%). Puede verse que esta tendencia refrenda la aparición de más sustantivos en la premodificación en inglés al menos en el discurso del genoma.

POS	Frecuencia	Porcentaje
N (sin núcleos)	51	45,95
Adj	36	32,43
PP	15	13,51
Adv	9	8,11

Tabla 2: Categoría léxica predominante en la premodificación

La explicación de esta preferencia por los sustantivos en la premodificación reside en el contenido del discurso. Al ser un corpus especializado, éste tiende a tener más sustantivos que adjetivos, puesto que la tendencia en un discurso más especializado en inglés apunta a que las secuencias de N concentran mayor densidad de nudos de conocimiento especializado. De igual modo, y como lo sugiere Halliday (1998: 193; Iturrioz 2000; Gallegos 2000, 2003), entre otros autores, existe una tendencia en el discurso científico a emplear nominalizaciones, con lo cual aumenta de entrada la cantidad de sustantivos que puede tener este tipo de discurso.

Un aspecto que este trabajo no ha explorado es la variación vertical (nivel de especialización) de estos datos y la variación horizontal (entre diferentes áreas del conocimiento), pues es posible que estos datos dependan no sólo del nivel de especialización (sintagmas nominales más extensos entre más especializado sea el texto), sino que varíen de área en área (áreas del conocimiento con diferentes niveles de abstracción, formas de comunicación, etc.), cuestión que sí se ha explorado con el corpus lexicográfico de contraste pues está conformado por diccionarios de varias áreas del conocimiento (Véase §4.4).

Es importante resaltar que aunque la cantidad de participios (13,51%) y adverbios (8,11%) no es alta a primera vista, estos porcentajes son altos si se comparan con estudios anteriores (Quiroz 2005) en los cuales, los participios no llegaban a 6,14% y 1,77%. Esto se debe, en primer lugar, a que los patrones extraídos contienen este tipo de categorías léxicas, en segundo lugar, a que el corpus es mayor en tamaño, lo que permite extraer patrones que son menos frecuentes en corpus de tamaño reducido. Además, si se tiene en cuenta que los patrones que albergan estas dos unidades léxicas son menos frecuentes en este estudio, estos porcentajes son aún más relevantes. Biber *et al* (1999: 589) encontraron que los participios son más comunes en el registro académico que en los otros registros, lo cual corrobora. Los adverbios son relativamente poco comunes comparados con los sustantivos y los adjetivos. Sin embargo, los participios y los adverbios en la muestra de este estudio representan un tercio y un cuarto de los adjetivos, con lo cual no puede decirse que sean poco comunes. Más adelante, se verá el papel que juegan estas dos categorías cuando funcionan como pares del tipo Adv PP, Adv Adj, etc. dentro del sintagma.

Si bien los tipos de participios (de presente y pasado) tienen funciones diferentes no sólo al nivel de la oración sino al nivel de la sintaxis del sintagma, no fue posible separarlos durante la extracción dado que los etiquetadores no los diferencian claramente o tienen problemas de desambiguación. Sin embargo, en el análisis semántico se podrá diferenciar su función y se asociarán

a sus respectivos patrones. La importancia y la función del participio de pasado, por ejemplo, como categoría premodificadora del sustantivo en el discurso especializado ya habían sido puestas de relieve por Swales (1985: 42-43). Para Swales (1985: 42) la posición pronominal de los participios de pasado puede estar asociada con rasgos generales o característicos y permanentes. De igual modo, la importancia de estas dos categorías ha empezado a cobrar vigencia en estudios contrastivos. Por ejemplo, los estudios llevados a cabo por Boughedaoui (1995, 1996, 1997, 1998, 2001), Maniez (2001) y Ormord (2001) sobre la coocurrencia de algunos patrones binarios dentro de la premodificación en inglés y las posibles traducciones al francés son de total relevancia para el par inglés-español. De igual modo, Vivanco (1994: 755-757) hace un pequeño análisis del inglés al español sobre los procesos de nominalización y las funciones de estas dos categorías dentro de los sintagmas nominales complejos. Concluye que aunque formalmente son formas no personales del verbo, sintáctica y funcionalmente han dejado de serlo ya que muchos de ellos pasan a ser verdaderos sustantivos y adjetivos (Vivanco 1994: 757).

4.3.3 Frecuencia de los patrones por aparición

La distribución de patrones por número de tokens de la tabla 3 muestra que los patrones más frecuentes son los de 3 tokens con una media de 75,75 sintagmas por patrón, luego siguen los patrones de 4 tokens con una media de 7,94 sintagmas por patrón y, por último, los patrones de 5 tokens con una media de 2,2 sintagmas por patrón. Estos datos muestran que la variabilidad en los patrones de superficie en los patrones menos extensos es menor si se compara con la alta variabilidad de los patrones más extensos que es de casi un patrón por cada dos sintagmas. Como se verá en §8, está variabilidad hace difícil la sistematización de este tipo de patrones no sólo para propósitos de traducción sino también para propósitos de terminología y enseñanza de lenguaje especializados. Y son precisamente estos los que más problemas de ambigüedad generan en el lector no experto, e. g. estudiante universitario, traductor, etc.

Igualmente, es necesario señalar que también los patrones más extensos (seis en adelante) son los menos frecuentes y que la probabilidad de encontrarlos en un texto es baja. De ahí que este estudio se centre en los más frecuentes que igualmente pueden generar problemas de interpretación en el hablante no por el aspecto conceptual sino por el aspecto lingüístico (Trimble 1985: 131).

Tokens	Patrones	Porcentaje	Frecuencia
3	12	86,16	909
4	17	12,8	135
5	5	1,04	11
Total	34	100	1.055

Tabla 3: Distribución de patrones por extensión en el corpus de análisis en inglés.

En la tabla 4 se presentan los 20 patrones más frecuentes de la muestra. Estos 20 patrones representan el 97,35% del total de la muestra con 1.027 ocurrencias sobre un total de 33 patrones y 1.055 ocurrencias. Es decir, que, en el resto de la muestra, existe una gran variabilidad en los 13 patrones restantes y 30 ocurrencias; es decir, hay casi un patrón por cada 3 sintagmas. En cambio, en estos 20 patrones existe una relación de 1 patrón por cada 51 sintagmas, lo que permite llevar a cabo generalizaciones más confiables, cuestión que no es muy factible con los otros 13 patrones.

De igual modo, puede verse que los primeros 10 patrones representan el 88,82% de toda la muestra con 937 sintagmas sobre los 90 sintagmas de los segundos 10 patrones (8,53%).

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N N N	polymerase chain reaction	317	30,05
3	Adj N N	horizontal gene transfer	254	24,08
3	Adj Adj N	human genomic DNA	113	10,71
3	N Adj N	platelet dense granules	62	5,88
3	PP N N	reduced insulin responsiveness	51	4,83
4	Adj N N N	fetal brain cDNA library	33	3,13
3	PP Adj N	polarized epithelial cells	31	2,94
4	N N N N	restriction fragment length polymorphism	28	2,65
3	Adv Adj N	anatomically modern humans	25	2,37
3	N PP N	ATP binding site	23	2,18
4	Adj Adj N N	human peripheral blood lymphocytes	16	1,52
3	Adv PP N	genetically engineered microorganisms	19	1,8
3	Adj PP N	neutral buffered formalin	13	1,23
4	N Adj N N	immunoglobulin heavy chain locus	10	0,95
4	Adj N Adj N	human APOE genomic DNA	9	0,85
4	PP N N N	pulsed field gel electrophoresis	6	0,57
4	PP Adj N N	inherited mitochondrial DNA diseases	5	0,47
4	Adj Adj Adj N	total human genomic DNA	4	0,38
4	Adv Adj N N	highly deleterious mtDNA mutations	4	0,38
4	Adv PP N N	highly conserved phosphotyrosine domain	4	0,38

Tabla 4: Los 20 patrones más frecuentes del corpus de análisis en inglés.

Entre los primeros 20 del corpus de análisis, los patrones de 3 y 4 tokens están repartidos al 50%. Sin embargo, su distribución en cuanto a la frecuencia es desigual, pues en los 10 primeros patrones de la tabla 4, 8 son de 3 tokens y tan sólo 2 patrones son de 4 tokens. Al contrario, entre los 10 últimos de la tabla, 8 patrones son de 4 tokens y 2 patrones de 3 tokens pero las ocurrencias son más bajas (937 vs. 90). Este aspecto indica de nuevo que existe una relación inversamente proporcional entre la extensión de un patrón y su aparición en la lengua. Es decir, entre menos extenso, más posibilidades tiene de aparecer y entre más extenso, menos posibilidades tiene de ocurrir. Por tanto, es probable que las unidades de menos extensión tiendan a lexicalizarse y es posible que los diccionarios especializados tengan más patrones o unidades de 3 tokens, cuestión que se abordará más adelante. Igualmente, estos datos muestran que la extensión está directamente relacionada con la estabilidad de dichas estructuras al ser más frecuentes y que una mayor variabilidad está directamente relacionada con una premodificación más extensa (Quiroz 2006: 380). En toda

la muestra, los 3 patrones más frecuentes son N N N, Adj N N y Adj Adj N. En conjunto, agrupan 684 sintagmas que representan un 64,84% del total de la muestra y, por extensión, del corpus de análisis. Los tres patrones son de 3 tokens de extensión y son los que tenderán más a lexicalizarse.

Si se analizan los patrones de acuerdo con el tipo de categoría léxica presente en la premodificación, puede verse que hay 14 patrones de los 20 más frecuentes que tienen uno o más sustantivos en la premodificación y equivalen al 77,92% (822 ocurrencias), como se presenta en la tabla 5. Esto demuestra que el sustantivo es la categoría léxica por elección en la premodificación al menos en este tipo de discurso. Se explica este gran número de patrones y ocurrencias porque son los sustantivos los que naturalmente tienden a representar objetos, procesos, fenómenos, etc. en el discurso científico-técnico y, los que además, nominalizan las acciones propias de los verbos. Puesto que una de las funciones de la premodificación es vehicular una gran cantidad de información en poco espacio de modo efectivo y eficiente, es el sustantivo la categoría prototípica para hacerlo en este tipo de discurso.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N N N[47]	polymerase chain reaction	317	30,05
3	Adj N N	horizontal gene transfer	254	24,08
3	N Adj N	platelet dense granules	62	5,88
3	PP N N	reduced insulin responsiveness	51	4,83
4	Adj N N N	fetal brain cDNA library	33	3,13
4	N N N N	restriction fragment length polymorphism	28	2,65
3	N PP N	ATP binding site	23	2,18
4	Adj Adj N N	human peripheral blood lymphocytes	16	1,52
4	N Adj N N	immunoglobulin heavy chain locus	10	0,95
4	Adj N Adj N	human APOE genomic DNA	9	0,85
4	PP N N N	pulsed field gel electrophoresis	6	0,57
4	PP Adj N N	inherited mitochondrial DNA diseases	5	0,47
4	Adv Adj N N	highly deleterious mtDNA mutations	4	0,38
4	Adv PP N N	highly conserved phosphotyrosine domain	4	0,38

Tabla 5: Patrones con uno o más sustantivos en la premodificación.

[47] Son precisamente este patrón y N N N N los más estudiados en inglés (ver Horsella y Pérez 1990).

En la tabla 6 puede observarse que hay 13 patrones de los 20 más frecuentes que tienen uno o más adjetivos en la premodificación y son el 54,89% (579 ocurrencias).

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	Adj N N	horizontal gene transfer	254	24,08
3	Adj Adj N	human genomic DNA	113	10,71
3	N Adj N	platelet dense granules	62	5,88
4	Adj N N N	fetal brain cDNA library	33	3,13
3	PP Adj N	polarized epithelial cells	31	2,94
3	Adv Adj N	anatomically modern humans	25	2,37
4	Adj Adj N N	human peripheral blood lymphocytes	16	1,52
3	Adj PP N	neutral buffered formalin	13	1,23
4	N Adj N N	immunoglobulin heavy chain locus	10	0,95
4	Adj N Adj N	human APOE genomic DNA	9	0,85
4	PP Adj N N	inherited mitochondrial DNA diseases	5	0,47
4	Adj Adj Adj N	total human genomic DNA	4	0,38
4	Adv Adj N N	highly deleterious mtDNA mutations	4	0,38

Tabla 6: Patrones que tienen uno o más adjetivos en la premodificación.

La función atributiva propia de adjetivo se ve reflejada en un buen número de patrones que tienen esta categoría léxica. Sin embargo, tienen un 23% menos de presencia en los patrones más frecuentes que los sustantivos.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	Adv Adj N	anatomically modern humans	25	2,37
3	Adv PP N	genetically engineered microorganisms	19	1,8
4	Adv Adj N N	highly deleterious mtDNA mutations	4	0,38
4	Adv PP N N	highly conserved phosphotyrosine domain	4	0,38

Tabla 7: Patrones con adverbios en la premodificación.

En la tabla 7 aparecen los 4 patrones de los 20 más frecuentes que tienen un adverbio en la premodificación y equivalen al 4,93% (52 ocurrencias). Puede observarse que el adverbio en todos los patrones se encuentra en la posición más lejana respecto del núcleo y está modificando a un adjetivo o un participio de pasado y estos en su conjunto al núcleo. La estructura que presenta más ocurrencias es Adv Adj N con 25 sintagmas, seguida por Adv PP N, ambos de 3

tokens de longitud. Puede considerarse que estas dos estructuras con adverbios son relativamente frecuentes si se tiene en cuenta que el adverbio es una de las categorías léxicas menos frecuentes en el sintagma.

En la tabla 8 aparecen los 8 patrones de los 20 más frecuentes que tienen sólo un participio en la premodificación y equivalen al 14,4% (152 ocurrencias). La mayoría de participios que se han detectado en las muestras son participios de pasado. Su función atributiva dentro de la premodificación es generalmente el resultado de la lexicalización de una oración pasiva, como los plantea Gotti (2003: 70): "The passive construction is also avoided by turning the verb into a past participle and using the latter as a premodifier", como puede verse en el sintagma del ejemplo 12.

12. reduced insulin responsiveness (responsiveness of the insulin which is reduced)

Este mismo hecho lo corrobora Boughedaoui (2001: 142) para el inglés y francés:

> Il s'agit de structures adjectivales où le deuxième élément est soit un participe passé, soit un participe présent. S'agissant du premier cas, le sens du participe passé employé comme épithète est en general passif, car cette structure émane de la transformation d'une proposition relative passive.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	PP N N	reduced insulin responsiveness	51	4,83
3	PP Adj N	polarized epithelial cells	31	2,94
3	N PP N	ATP binding site	23	2,18
3	Adv PP N	genetically engineered microorganisms	19	1,8
3	Adj PP N	neutral buffered formalin	13	1,23
4	PP N N N	pulsed field gel electrophoresis	6	0,57
4	PP Adj N N	inherited mitochondrial DNA diseases	5	0,47
4	Adv PP N N	highly conserved phosphotyrosine domain	4	0,38

Tabla 8: Patrones con participios en la premodificación.

El participio de pasado suele estar a su vez modificado por un adverbio, especialmente terminado en el sufijo *-ly* (*-mente*) como resultado de una lexicalización de una oración en voz pasiva modificada por un adverbio, como se ilustra en los ejemplos de 13 y 14.

13. genetically engineered microorganisms (microorganisms which are genetically engineered)
14. polarized epithelial cells (epithelial cells which are polarized)

Puesto que el comportamiento de los participios de presente es flexible, no es posible recuperar su origen en muchos casos. Sager *et al* afirman (1980: 215-217) que las formas terminadas en *-ing* pueden funcionar como adjetivos, participios y sustantivos dentro de la misma premodificación, lo que lo hace una categoría muy versátil dentro de la comunicación especializada para denominar procesos y métodos. Boughedaoui (2001: 143) explica que la lexicalización de los participios de presente en sustantivos resulta de un verbo en voz activa, como se observa en los ejemplos 15 y 16.

15. ATP binding site (a site which binds to the ATP)
16. gene mapping studies (studies that map genes)

Para observar el predominio de una u otra categoría léxica, se han separado los patrones que contienen sólo adjetivos o sustantivos en los 20 patrones más frecuentes. En tabla 9 se presentan los 6 patrones que no tienen sustantivo en la premodificación y equivalen al 19,43% (205 ocurrencias) del total de la muestra. Obsérvese que sólo un patrón carece de adjetivo (Adv PP N).

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	Adj Adj N	human genomic DNA	113	10,71
3	PP Adj N	polarized epithelial cells	31	2,94
3	Adv Adj N	anatomically modern humans	25	2,37
3	Adv PP N	genetically engineered microorganisms	19	1,8
3	Adj PP N	neutral buffered formalin	13	1,23
4	Adj Adj Adj N	total human genomic DNA	4	0,38

Tabla 9: Patrones sin sustantivos en la premodificación.

En la tabla 10 los 7 patrones que no tienen adjetivo en la premodificación equivalen al 42,46% (448 ocurrencias), pero tienen sustantivo en su mayoría, excepto el patrón Adv PP N, de nuevo. Esto refrenda las observaciones hechas antes no sólo en cuanto a que el sustantivo es la categoría léxica que más predomina en los patrones, sino que hay más patrones con sustantivos (tabla 5) que no tienen adjetivo. Existe más del doble de patrones que no tienen adjetivos que aquellos que no tienen sustantivos, lo que demuestra la preferencia del discurso científico-técnico por las nominalizaciones.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N N N	polymerase chain reaction	317	30,05
3	PP N N	reduced insulin responsiveness	51	4,83
4	N N N N	restriction fragment length polymorphism	28	2,65
3	N PP N	ATP binding site	23	2,18
3	Adv PP N	genetically engineered microorganisms	19	1,8
4	PP N N N	pulsed field gel electrophoresis	6	0,57
4	Adv PP N N	highly conserved phosphotyrosine domain	4	0,38

Tabla 10: Patrones sin adjetivos en la premodificación.

Finalmente, en la tabla 11 se presentan los 8 patrones de los 20 más frecuentes que tienen tanto adjetivo como sustantivo en la premodificación y equivalen al 37,26% (393 ocurrencias). De ellos, 2 patrones son de 3 tokens y 6 de 4 tokens.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	Adj N N	horizontal gene transfer	254	24,08
3	N Adj N	platelet dense granules	62	5,88
4	Adj N N N	fetal brain cDNA library	33	3,13
4	Adj Adj N N	human peripheral blood lymphocytes	16	1,52
4	N Adj N N	immunoglobulin heavy chain locus	10	0,95
4	Adj N Adj N	human APOE genomic DNA	9	0,85
4	PP Adj N N	inherited mitochondrial DNA diseases	5	0,47
4	Adv Adj N N	highly deleterious mtDNA mutations	4	0,38

Tabla 11: Patrones con adjetivos y sustantivos en la premodificación.

4.3.4 Frecuencia de los patrones por longitud

Los patrones más frecuentes distribuidos por la cantidad de tokens se presentan a continuación de mayor a menor extensión.

En la tabla 12 los datos muestran que debido a la variabilidad estructural de este tipo de sintagmas, no existe un patrón que sea más productivo que permita obtener regularidades a este nivel. Sin embargo, puede observarse en los ejemplos que todos son unidades especializadas y que intuitivamente forman unidades de conocimiento. Estas unidades funcionan como modos de expansiones de unidades de menos tokens. Por ejemplo, el patrón N N N N N es una expansión del patrón N N N N que es el segundo patrón más extenso de los patrones de 4 tokens con una frecuencia media, y este a su vez del patrón N N N que es el más frecuente de todos los patrones de este estudio. En el ejemplo 17, se presenta la expansión del patrón N N N y sus ocurrencias en el buscador Google. Obsérvese en el ejemplo 17 que la frecuencia disminuye a medida que aumenta la extensión y la especialización del sintagma.

17. V1aR mRNA transcription start site (10 veces)[48]
mRNA transcription start site (333 veces)
transcription start site (695.000 veces)

Patrón	Ejemplo	Frecuencia	Porcentaje
Adj Adj N N N	human mitochondrial half ABC protein	3	0,28
Adj N N N N	Human Prostaglandin F Receptor Gene	3	0,28
N N N N N	V1aR mRNA transcription start site	3	0,28
Adj Adj Adj N N	mature neuronal nicotinic acetylcholine receptors	1	0,09
Adv PP N N N	covalently closed plasmid DNA band	1	0,09

Tabla 12: Los patrones más frecuentes de 5 tokens en el corpus de análisis en inglés.

[48] Consultado el 12 de marzo de 2007 en el buscador Google.

En la tabla 13 se presentan los patrones de 4 tokens (con el núcleo). En total, la muestra contiene 17 patrones de 4 tokens que representan un 12,8% y 135 ocurrencias. Los patrones de 4 tokens agrupan casi la mitad de los patrones de la muestra y, hasta cierto punto, presentan una variabilidad sintáctica importante al tener una relación de 1 patrón por cada 8 sintagmas. El patrón más frecuente es Adj N N N con 3,13% y 33 ocurrencias, seguido muy de cerca por el patrón N N N N con un 2,65% y 28 ocurrencias.

Puede observarse que, a diferencia de los patrones de 5 tokens, hay patrones que tienen una frecuencia mayor que otros como sucede con los 5 primeros de la tabla, los cuales representan más de la mitad de las ocurrencias de los otros 12 patrones (96 contra 39) y tienen una alta frecuencia en el corpus.

Patrón	Ejemplo	Frecuencia	Porcentaje
Adj N N N	fetal brain cDNA library	33	3,13
N N N N	restriction fragment length polymorphism	28	2,65
Adj Adj N N	human peripheral blood lymphocytes	16	1,52
N Adj N N	ELT-3 smooth muscle cells	10	0,95
Adj N Adj N	high affinity human antibodies	9	0,85
PP N N N	pulsed field gel electrophoresis	6	0,57
PP Adj N N	increased fat cell number	5	0,47
Adj Adj Adj N	cytogenetic bacterial artificial chromosome	4	0,38
Adv Adj N N	highly deleterious mtDNA mutations	4	0,38
Adv PP N N	locally produced growth factors	4	0,38
Adv Adj Adj N	morphologically identifiable apoptotic cells	3	0,28
N N PP N	Arabidopsis suspension cultured cell	3	0,28
PP Adj Adj N	polarized renal epithelial cells	3	0,28
Adv PP Adj N	darkly stained apical dendrites	2	0,19
N Adj Adj N	rabbit fast skeletal muscle	2	0,19
N PP N N	suspension- cultured Arabidopsis cells	2	0,19
N Adv N N	fluorescence in situ hybridization probes	1	0,09

Tabla 13: Los patrones más frecuentes de 4 tokens en el corpus de análisis en inglés.

En la tabla 14 se presentan los patrones de 3 tokens. Estos representan un 86,16% de toda la muestra con 909 ocurrencias. Existe una variabilidad sintáctica de 1 patrón por cada 75,75 sintagmas. Salvo por el patrón PP PP N, todos los patrones de 3 tokens tienen una alta frecuencia, lo que puede revertir en estructuras estables y con tendencia a que estas unidades sean términos.

Patrón	Ejemplo	Frecuencia	Porcentaje
N N N	polymerase chain reaction	317	30,05
Adj N N	horizontal gene transfer	254	24,08
Adj Adj N	human genomic DNA	113	10,71
N Adj N	yeast artificial chromosome	62	5,88
PP N N	reduced insulin responsiveness	51	4,83
PP Adj N	polarized epithelial cells	31	2,94
Adv Adj N	environmentally dependent phenotype	25	2,37
N PP N	nucleotide binding pocket	23	2,18
Adv PP N	genetically engineered microorganisms	19	1,8
Adj PP N	neutral buffered formalin	13	1,23
PP PP N	written informed consent	1	0,09

Tabla 14: Los patrones más frecuentes de 3 tokens en el corpus de análisis en inglés.

El patrón más frecuente es N N N con un 30% del total de la muestra y 317 sintagmas, seguido por el patrón Adj N N con un 24,08% y 254 sintagmas. A su vez son las estructuras más frecuentes en todo el corpus de análisis en inglés.

Una vez más, se observa que la longitud de los patrones incide directamente en la variabilidad sintáctica, es decir, a mayor extensión, más variabilidad existe y a menor extensión, menor variabilidad.

De igual modo, la productividad de los patrones tiende a disminuir con la extensión. Entre más extenso sea un patrón, menos productivo será y entre menos extenso sea, más productivo será. De hecho, puede considerarse que los patrones de productividad media a baja de 3 tokens (4,86% y un rango entre 30,05% y 1,23%) son más productivos que el patrón más productivo de 4 tokens (0,28%) y de 5 tokens (3,13%). Sin embargo, esta productividad también puede ser un problema desde el punto de vista terminológico ya que los patrones más productivos pueden contener muchas unidades que no son terminológicas, mientras que aquellos de productividad media pueden presentar menos ruido. Este aspecto podrá observarse más adelante con el corpus lexicográfico (§4.4).

4.3.5 Relaciones de dependencia del corpus de análisis en inglés

Como se explicó en el capítulo de la metodología (§3.7), para el análisis de dependencias sintácticas de los patrones en inglés, se seleccionó manualmente una muestra de los 10 patrones más frecuentes a partir de la muestra del análisis morfosintáctico. Estos 10 patrones representan 88,82% de toda la muestra con 937 sintagmas. Para ello, se seleccionó un 24,37% de los sintagmas y se distribuyó proporcionalmente de acuerdo con su frecuencia, como se hizo con la muestra sintáctica. Es decir, al patrón más frecuente, le corresponden más sintagmas para el análisis semántico y al patrón menos frecuente se le asignan menos sintagmas. Por ejemplo, el patrón N N N es el más frecuente del corpus y le corresponden 54 sintagmas y el menos frecuente es el Adv PP N N y le corresponden 4 sintagmas. A su vez esta muestra se empleó para el análisis semántico de §6.

En la tabla 15 se presenta la frecuencia de dependencias en el conjunto de patrones en inglés. La relación de dependencia [C [[B A]] es la más frecuente en todo el corpus en inglés con más del 61,2% de todas las ocurrencias (142) del corpus de análisis.

Dependencia	Frecuencia	Porcentaje
[C [[B A]]	142	61,2
[[C B] A]	67	28,89
[[D C] [B A]]	12	5,17
Ambigua	11	4,74

Tabla 15: Frecuencias de las dependencias de los patrones en inglés

En esta dependencia el primer modificador forma un conjunto con el núcleo del sintagma a manera de compuesto sintagmático y el modificador externo lo modifica, como se ejemplifica en 18.

18. metastatic colorectal cancer, columnar epithelial cells, central nervous system, white blood cells, red blood cells, aberrant FHIT transcript, male sexual

orientation, backbone nuclear resonances, helper T cells, mouse Igf2r gene, DNA Sequencing Kit, TA Cloning Kit, expected molecular mass, circulating monoclonal protein, polarized epithelial cells, pulverized rat chow, reduced insulin responsiveness

Posteriormente, sigue la relación de dependencias [[C B] A] con un 28,89% de todas las ocurrencias (67). En esta dependencia, los dos premodificadores forman un conjunto para modificar al núcleo, como se muestra en los ejemplos de 19.

19. mitochondrial DNA mutations, smooth muscle cells, mitochondrial DNA molecules, adipose cell size, fetal brain library, eukaryotic DNA metabolism, highly polymorphic markers, environmentally dependent phenotype, highly polymorphic markers, bone marrow aspirate, FHIT gene transcript, T cell responses, bile duct ligation, carbon tetrachloride model, Hepatitis B virus

Por último, aparece la dependencia [[D C] [B A]] con un 5,17% de todas las ocurrencias (12) para patrones de 4 tokens, como se observa en los ejemplos de 20.

20. somatic cell hybrid panel, fetal brain cDNA library, somatic cell hybrid analysis, somatic cell hybrid DNA, white blood corpuscle count, American Type Culture Collection, antibiotic resistance marker genes, plasmid DNA production process, Protein A affinity chromatography, potassium channel gene cluster

En la tabla 16, se presentan las relaciones de dependencia de cada uno de los patrones.

El patrón Adj Adj N tiene todas sus ocurrencias (28) con la dependencia [C [[B A]]. Por ejemplo, en el sintagma *human peripheral blood*, el primer adjetivo *peripheral* modifica directamente al núcleo *blood* para formar un conjunto *peripheral blood* y, luego el segundo adjetivo *human* modifica al conjunto *peripheral blood*, como puede también observarse en los casos de 21.

21. green fluorescent protein, human peripheral blood, total cellular RNA, central nervous system, columnar epithelial cells, fetal bovine serum, human fetal brain, human genomic DNA, human genomic library, human mitochondrial DNA, human placental DNA, immature leukocytic cells, immunoreactive glial cells, inner nuclear membrane, large human chromosome, metastatic colorectal cancer, multiple congenital abnormalities, normal human cortex, total genomic DNA, total human DNA

Patrón	Dependencia	Frecuencia	Porcentaje
Adj Adj N	[C [[B A]]	28	12,06
Adj N N	[C [[B A]]	46	19,82
Adj N N	[[C B] A]	22	9,48
Adj N N	Ambiguo	1	0,431
Adj N N N	[[D C] [B A]]	7	3,017
Adv Adj N	[[C B] A]	6	2,586
N Adj N	[C [[B A]]	14	6,034
N Adj N	Ambiguo	1	0,431
N N N	[[C B] A]	39	16,81
N N N	[C [[B A]]	28	12,07
N N N	Ambiguo	8	3,448
N N N N	[[D C] [B A]]	5	2,155
N N N N	Ambiguo	1	0,431
N PP N	[C [[B A]]	6	2,586
PP Adj N	[C [[B A]]	9	3,879
PP N N	[C [[B A]]	11	4,741

Tabla 16: Dependencias de los patrones de la muestra de análisis en inglés.

El patrón Adj N N tiene dos relaciones de dependencia [C [[B A]] con 46 ocurrencias (64,78%) y [[C B] A] con 22 (35,21%) de un total de 71 ocurrencias. En el caso de la dependencia [C [[B A]], que representa a la mayoría de ocurrencias, el sustantivo premodificador *blood* forma un tipo de compuesto con el núcleo *blood cell*, este conjunto es modificado por el adjetivo *red* para formar el sintagma *red blood cell* y los sintagmas de 22.

22. horizontal gene transfer, human NGF gene, white blood cells, prandial insulin infusion, human X chromosome, apoptotic cell death, Human IL11RA gene, outer root sheath, embryonic stem cells, endothelial growth factor, human IGF2R Gene, human MRP genes, human tau gene, human TnTf gene, inner root sheath, epithelial root sheath, epidermal growth factor, human HMGIC gene,

aberrant FHIT transcript, human APOE gene, human YAC library, individual TLE genes, basolateral cell membrane

En cambio, en la dependencia menos frecuente para el patrón Adj N N, [[C B] A], el adjetivo *adipose* modifica al sustantivo premodificador *cell* y este conjunto *adipose cell* al núcleo *size* para formar el sintagma *adipose cell size* y los casos que se presentan en 23.

23. chemical shift changes, mitochondrial DNA mutations, genomic DNA fragments, smooth muscle cells, cytoplasmic membrane protein, bovine serum albumin, mitochondrial DNA molecules, fetal calf serum, genomic DNA clones, human metaphase chromosomes, meiotic recombination distance, Southern blot hybridization, adipose cell size, chemical shift differences, genomic DNA fragment, contiguous gene syndromes, fetal brain library, eukaryotic DNA metabolism

El patrón Adj N N N tiene todas las ocurrencias (7) regidas por la dependencia [[D C] [B A]], como se ejemplifica en 24. En este caso el sintagma el adjetivo *fetal* modifica *brain* y el sustantivo *cDNA* al núcleo *library*. Posteriormente el sintagma *somatic cell* modifica al sintagma *cDNA library* para formar el sintagma *fetal brain cDNA library*.

24. somatic cell hybrid panel, fetal brain cDNA library, somatic cell hybrid analysis, unequal variance t test, somatic cell hybrid DNA, white blood corpuscle count, American Type Culture Collection

Al igual que el patrón anterior, el patrón Adv Adj N está regido por la misma dependencia [[C B] A] con 6 ocurrencias, como se observa en los ejemplos de 25. En el sintagma *highly polymorphic markers*, el adverbio *highly* modifica al adjetivo *polymorphic* para formar el sintagma adjetival *highly polymorphic* para modificar conjuntamente al núcleo *markers*.

25. anatomically modern humans, right ventricular myocardium, slightly deleterious mutations, environmentally dependent phenotype, highly polymorphic markers, highly epistatic genes

El patrón N Adj N tiene también casi la totalidad de las ocurrencias regidas por la dependencia [C [[B A]], como se ve en los ejemplos de 26.

26. platelet dense granules, yeast artificial chromosome, glucose specific activity, BAC genomic clone, male sexual orientation, myosin heavy chain, MUC7 genomic clones, Genius nonradioactive DNA, backbone nuclear resonances, herpes simplex virus, lung lysosomal enzymes, kidney lysosomal enzymes, APOE Genomic DNA, apoE neuronal immunoreactivity, APOE transgenic mice

El patrón N N N tiene básicamente 2 formas de dependencia: [[C B] A] con 39 ocurrencias (87,93%) y [C [[B A]] con 28 ocurrencias (12,07%). En la primera dependencia [[C B] A], los sustantivos de la premodificación *potassium channel* forman un sintagma nominal que modifica directamente al núcleo *gene* para formar el sintagma *potassium channel gene*, como sucede con los casos de 27.

27. ELT-3 cell growth, LIM domain proteins, metaphase chromosome spreads, amyloid subunit protein, amino acid changes, K2 cell monolayers, plasmid DNA production, tyrosine kinase activity, amino acid identity, amino acid level, animal cell pol, T cell responses, ArG promoter polymorphism, cytokine gene polymorphism, LIM domain protein, plasmid copy number, plasmid DNA vaccines, calcium phosphate method, amino acid differences, antibiotic resistance marker, carbon tetrachloride model, plasmid copy number, amino acid sequence, bone marrow aspirate, carbon tetrachloride model, stellate cell activation, FHIT gene transcript, FHIT transcript aberration, APOE knockout mice, APOE knockout mouse, CTD phosphate turnover

Sin embargo, bajo esta misma dependencia se encuentran ejemplos como el sintagma *hepatitis B virus*, en el cual, el núcleo del sintagma nominal premodificador, *hepatitis B* no es el sustantivo B sino el sustantivo *hepatitis*. En

este caso el que hace las veces de modificador es el sustantivo *B*. Luego este sintagma premodifica en su conjunto al núcleo *virus*. Aunque es un solo caso de la muestra, existen otros casos en el corpus, y en general, todos aquellos que tienen un sustantivo con carácter nomenclador (e.g *type I collagen*).

El patrón N N N N tiene una sola forma de dependencia: [[D C] [B A]] con 5 ocurrencias (83,33%) aunque hay 1 ambiguo. En esta dependencia, el sustantivo *plasmid* modifica a *DNA* y el sustantivo *production* al núcleo *process*. Luego el sintagma nominal *somatic cell* modifica al sintagma nominal *cDNA library* para formar el sintagma *plasmid DNA production process*, como sucede con los casos de 27 presentados antes.

El patrón N PP N tiene una sola forma de dependencia: [C [[B A]] con 6 ocurrencias. En este patrón, el participio *sequencing* modifica al núcleo *kit* y estos son modificados por el sustantivo *DNA* para formar el sintagma *DNA sequencing kit*, como se ve en los ejemplos de 28.

28. nucleotide binding pocket, SDS loading buffer, mutant processing proteins, T7 sequencing kit, DNA sequencing kit, TA cloning kit

El patrón PP Adj N tiene una sola forma de dependencia: [C [[B A]] con 9 ocurrencias. En este patrón, el adjetivo *epithelial* modifica al núcleo *cells* y estos son modificados por el participio *polarized* para formar el sintagma *polarized epithelial cells*, como se ilustra en los ejemplos de 29.

29. polarized epithelial cells, circulating monoclonal protein, expected molecular mass, increased chromosomal breakage, automated thermal cycler, ragged red fibers, repeated auditory stimuli, repressed paternal allele, biotinylated genomic fragment

El patrón PP N N tiene una sola forma de dependencia: [C [[B A]] con 11 ocurrencias. En este patrón, *HIV-1* modifica al núcleo *disease* y estos son

modificados por el participio *advanced* para formar el sintagma *advanced HIV-1 disease*, como se ve en los ejemplos de 30.

30. growing polypeptide chain, reduced insulin responsiveness, masked study medication, advanced HIV-1 disease, expected PCR product, increased chromosome breakage, known HLA-DR2 association, striated muscle contraction, published cDNA sequence, pulverized rat chow, verified mutation carriers

Hay 11 sintagmas que se han etiquetado como ambiguos dado que sería necesario el conocimiento de un especialista o experto en el ámbito para poder asignar la dependencia o en algunos casos dicha dependencia no es del todo clara como sucede con el caso de *HBV DNA polymerase*, en el cual *HBV DNA* (hepatitis B virus DNA) existe como término según el glosario de *HB Foundation*[49] y *DNA polymerase* como enzima según el diccionario Stedman de medicina.

> HBV DNA is a marker of viral replication and level of infectivity. It is used to assess and monitor the treatment of patients with chronic HBV infection.

> DNA polymerase -> nucleotidyltransferases: enzymes (EC class 2.7.7) transferring nucleotide residues (nucleotidyls) from nucleoside di- or triphosphates into dimer or polymer forms. Some nucleotidyltransferases's bear specific names (e.g., adenylyltransferases), or trivial names indicating the linkage hydrolyzed in the synthesis (pyrophosphorylases, phosphorylases), or names of the material synthesized (RNA or DNA polymerase).

[49] www.hepb.org/expforum/glossary.aspx

4.4 Resultados del corpus lexicográfico de contraste en inglés

No cabe duda de que la terminología es un elemento clave de los textos y los discursos especializados, como han señalado muchos autores (Sager 1990; Kocourek 1991; Cabré 1992, 1999, 2002) y, por lo tanto, también lo es para la traducción especializada (Maillot 1981; Bédard 1986; Durieux 1988; Hans 1990; Sager 1992; Wright 1993; Durieux 1997; Scarpa 2001).

En consecuencia, es lógico que la resolución de problemas terminológicos ocupe una buena parte del esfuerzo que un traductor dedica a la traducción de un texto especializado (Scarpa 2001: 154). Algunos autores (Arntz 1993: 5; Walker 1993: 22; Sager 1990, 1994: 206, Fähndrich 2005: 239) han estimado que cerca del 40% ó 50% del tiempo invertido durante el proceso de una traducción especializada se dedica a la resolución de problemas terminológicos, y subrayan que una gestión adecuada de la terminología en el proceso de traducción asegura no solamente unos niveles más altos de calidad en el texto, sino de productividad en el ejercicio de su trabajo. Estos datos se corroboran en la reciente encuesta *Terminology Survey 2005* (Zielenski, Ramírez 2005: 19) en la que los traductores estiman que el tiempo total empleado en la gestión terminológica en un encargo de traducción es del 30% y consideran que en traducciones muy especializadas dicho porcentaje puede alcanzar hasta un 70% del tiempo total, lo que concuerda con lo expuesto por Scarpa (2001: 154):

> Uno degli aspetti fondamentali che caratterizzano l'attività del traduttore è la ricerca delle corrispondenze terminologico-concettuali nelle lingue di partenza e di arrivo, che diventa tanto più laboriosa quanto maggiore è il livello di specializzazione del testo da tradurre.

Como lo revela *Translation Memory Survey 2007* (13, 14), cuando los traductores profesionales se encuentran en el escenario de no saber la traducción de una oración en la lengua de llegada, el 83% de los traductores

recurren a los diccionarios como primera fuente de búsqueda bien sea en forma de diccionarios o glosarios en CD-ROM (30%), en diccionarios o glosarios en Internet (21%), en diccionarios o glosarios y en papel (17%) o en un buscador tipo Google (15%). En consecuencia, los diccionarios especializados son la principal y más confiable fuente de consulta de terminólogos, traductores, estudiantes universitarios y profesionales de las lenguas, ya que proporcionan diversos tipos de información que pueden ser útiles para los diferentes usuarios.

En la tabla 17 se presenta cada uno de los diccionarios consultados, su área temática con su número de entradas y la cantidad de términos de más de tres tokens de longitud y el porcentaje de esta última cantidad. Puede observarse que existe una tendencia que entre más extensos sean los diccionarios, menos cantidad de sintagmas de más de tres tokens tienen. Esto quiere decir que entre más grande sea un diccionario, más cantidad de unidades simples tiene. Por ejemplo, el diccionario Routledge de economía, que tiene unas 38.000 entradas, sólo tiene 5.269 sintagmas de más de tres tokens (13,86%) y 11.890 unidades simples (31,28 %). Si se revisara rápidamente algunas de estas unidades simples, puede verse que tienen un carácter terminológico dudoso o sólo funcionan como unidades terminológicas generales dentro del diccionario (p. ej., *omnibus, Web, above, able, have, hysteresis*, etc.).

Al contrario, entre menos entradas tenga un diccionario, más unidades de tres tokens de longitud tienen. Por ejemplo, el diccionario ISI de estadística tiene 4.500 entradas, de las cuales 1.238 son más de tres tokens (35,37%).

Este mismo hecho puede observarse en un estudio sobre la alineación de términos multipalabra (Daille *et al* 2004: 921). Estos autores tomaron los datos de tres repositorios terminológicos del inglés al francés en el área de la silvicultura y puede inferirse la misma relación entre el tamaño del diccionario de la cantidad de unidades multipalabra (700 términos y 70% de unidades multipalabra, 2.800 términos y 66% y 15.000 términos y 47%).

Puede verse además que las diferencias entre los diccionarios no radican en el área temática sino en el tamaño del diccionario. Al menos en este sentido, no existe una diferencia importante, es decir, no existe una relación entre el tema y la extensión de los sintagmas. Antes, se planteó que entre más extensos fueran los sintagmas, más especializados y menos estables tendían a ser. Por tanto, el grado de especialización de un diccionario no está regido en apariencia ni por la extensión de los sintagmas ni por el área temática, a pesar de que existe la idea generalizada de que este tipo de sintagmas es típico de ciertas áreas del conocimiento como lo propone Sager (1980: 272): "*... in practice five- or – six-element compounds are rare, but variations exist between special subjects*".

Diccionario	Área temática	N.º de entradas	SN de +3 tokens	Porcentaje
Diccionario Mosby	Medicina	31.400	3.553	11,31%
Diccionario IFCC	Lab. Clínico	4.039	725	17,94%
IMF Terminology	Economía	4.500	766	17,02%
Routledge Dictionary	Finanzas	38.000	5.269	13,86%
ISI Multilingual Glossary	Estadística	3.500	1.238	35,37%

Tabla 17: Datos de referencia de cada diccionario en inglés.

A continuación, se describen los resultados del corpus lexicográfico de contraste que se emplea para poder observar las tendencias de extensión, frecuencia de los patrones de más de 3 tokens en diferentes áreas del conocimiento.

4.4.1 Longitud y frecuencia de los SN en los diccionarios en inglés

En la tabla 18 se presentan los resultados de la longitud de los sintagmas en los diccionarios ordenados de menor a mayor (de tres tokens a siete tokens). Así, existe una relación directa entre la extensión del sintagma y la frecuencia de aparición en todos los diccionarios.

Diccionario	3 tokens	Porc.	4 tokens	Porc.	5 tokens	Porc.	6 tokens	Porc.	7 tokens	Porc.
Mosby	3006	84,6	437	12,3	87	2,45	17	0,048	6	0,017
IFCC	505	69,66	161	22,21	37	5,1	20	2,76	2	0,28
IMF	626	81,72	118	15,4	22	2,87				
Routledge	4446	84,38	695	13,19	106	2,01	18	0,034	4	0,008
ISI	978	79	221	17,85	34	2,75	4	0,032	1	0,008
Promedio		79,87		16,19		3,03		0,057		0,06

Tabla 18: Frecuencia por número de tokens del corpus lexicográfico de contraste en inglés.

En todo el corpus lexicográfico de contraste, los sintagmas de 3 tokens son los más frecuentes (9.561 ocurrencias y un 79,87% en promedio), como se ilustra en la tabla 20. Por el contrario, los sintagmas de más de 7 tokens son los menos frecuentes (13 ocurrencias y un 0,06% en promedio). Además, puede verse que los sintagmas de 3 y 4 tokens agrupan el 96,06% de todos los sintagmas del corpus lexicográfico, lo que una vez más confirma los resultados obtenidos por Cartagena (1998) y Quiroz (2005), en cuanto a que la extensión de los sintagmas está en el rango de 3 y 4 tokens.

En este corpus lexicográfico, solo el 3,94% representa al resto de sintagmas (de 5 a 8 tokens). Desde un punto de vista de la traducción, las unidades de más de 5 tokens son las que ofrecen más problemas a la hora de acuñar un equivalente en español y, por tanto, su bajo nivel de aparición en los diccionarios es una desventaja para el traductor ya que serían potencialmente unidades que buscaría durante el proceso de traducción.

4.4.2 Categoría léxica predominante en la premodificación de los SN en los diccionarios en inglés

En las gramáticas más importantes del inglés, (e.g. Biber *et al* 1999: 589) y libros de inglés para propósitos académicos (IAP), se establece que la categoría léxica más común en la premodificación es el adjetivo y no el sustantivo. Sin embargo, este estudio también confirma las observaciones hechas antes en Quiroz (2005) de que en el discurso especializado los sustantivos son más

frecuentes en la premodificación que los adjetivos. Esto se debe a que el discurso especializado emplea la nominalización como una estrategia discursiva para expresar impersonalidad y objetividad del discurso. Puesto que se deben usar muchos objetos, procesos y acciones para representar el conocimiento de un área, la premodificación es una forma efectiva de juntar sustantivos y reducir las oraciones.

Puede observarse en la tabla 19 que, salvo en el diccionario ISI, la categoría léxica predominante es el sustantivo. En casi todos los casos, los sustantivos casi duplican a los adjetivos con una media de 44,15% (rango entre 35,92% y 56,19%) mientras que la media de los adjetivos no supera el 25,02% (rango entre 16,31% y 37,14%). A continuación, siguen los participios de pasado con 7,47% (rango entre 3,32% y 11,43%), los numerales con un 5,47% (rango entre 2,04% y 5,64%), los participios de presente con 4,59% (rango entre 4,26% y 5,64%) y los adverbios con 6,95% (rango entre 2,42% y 9,64).

También se encuentran otras categorías léxicas como preposiciones, conjunciones, verbos, determinantes, prefijos y pronombres que en conjunto representan un 6,18%.

	Mosby		IFCC		IMF		Routledge		ISI	
POS	Frec.	Porcent.	Frec.	Porcent.	Frec.	Porcent.	Frec.	Porcent.	Frec.	Porcent.
N	210	43,84	186	56,19	84	42,64	208	42,19	88	35,92
Adj	123	25,68	54	16,31	47	23,86	109	22,11	91	37,14
PP	35	7,31	11	3,32	17	8,63	37	7,51	28	11,43
Num	30	6,26	40	12,08	5	2,54	22	4,46	5	2,04
PPi	27	5,64	16	4,83	9	4,57	21	4,26	9	3,67
Adv	25	5,22	8	2,42	19	9,64	40	8,11	23	9,39
Prep	19	3,97	8	2,42	10	5,08	37	7,51	1	0,41
Conj	7	1,46	3	0,91	0	0	5	1,01	0	0
V	2	0,42	2	0,6	3	1,52	13	2,64	0	0
Det	1	0,21	0	0	2	1,02	1	0,2	0	0
Prefix	0	0	3	0,91	0	0	0	0	0	0
Pron	0	0	0	0	1	0,51	0	0	0	0

Tabla 19: Categoría léxica predominante en la premodificación del corpus lexicográfico en inglés.

En cuanto al predominio de la categoría gramatical dentro de la premodificación del corpus lexicográfico, hay 52 patrones sin sustantivos en la premodificación de los 283 totales y 136 patrones sin adjetivos, lo que muestra el predominio de los sustantivos como categoría premodificadora. Por otro lado, hay 21 patrones que carecen de sustantivos y adjetivos y 116 patrones con sustantivos y adjetivos a la vez. En cuanto a las otras categorías léxicas abiertas, hay 60 patrones con adverbios, 61 patrones con participio de pasado y 41 con participio de presente.

4.4.3 Frecuencia de los patrones por aparición en inglés

En este ítem se presentan los resultados del corpus lexicográfico de contraste de acuerdo con su frecuencia en todo el corpus y en cada diccionario y también se presentan los datos de acuerdo con la extensión del sintagma en cada diccionario.

Diccionario	N.º de entradas	N.º de patrones	SN de +3 tokens en inglés	Promedio por patrón
Diccionario Mosby	31.400	143	3.553	24,84
Diccionario IFCC	4.039	94	725	7,71
IMF Terminology	4.500	66	766	11,6
Routledge Dictionary	38.000	157	5.269	33,56
ISI Multilingual Glossary	3.500	78	1.238	15,87
Total		538 (283 diferentes)	11.551	

Tabla 20: Número de patrones totales del corpus lexicográfico en inglés por diccionario y promedio por patrón.

En su conjunto, el corpus lexicográfico de contraste del inglés contiene 283 patrones diferentes. El diccionario con más patrones es el Routledge con 157 y un promedio de 33,56 sintagmas por patrón y el diccionario con menos patrones es el diccionario ISI con 78 patrones y una media de 15,87 sintagmas por patrón. Puede apreciarse en la tabla 20 que los diccionarios con mayor número de entradas tienen mayor variabilidad en cuanto a la cantidad de

patrones a pesar de que la relación del total de patrones contra el total de sintagmas de más de tres tokens es alta.

Sin embargo, como puede apreciarse en la tabla 21, los cinco primeros patrones de cada diccionario representan a la mayoría de ocurrencias (rango entre 68% y 77%), mientras que al resto de ocurrencias le corresponde un número importante de estructuras. Esto demuestra que también en el corpus lexicográfico de contraste existe una variabilidad sintáctica considerable.

Una vez más, estos datos muestran que la longitud de un sintagma está directamente relacionada con la estabilidad de las estructuras y que hay unas cuantas estructuras (6 ó 7) que representan a una gran cantidad de sintagmas. Igualmente, una mayor variabilidad sintáctica está relacionada directamente con una premodificación más extensa.

Como se ilustra en la tabla 21, los patrones más frecuentes en el corpus lexicográfico de contraste son: N N N que es el más frecuente en cuatro diccionarios y es el segundo más frecuente en uno de ellos; Adj N N que es el segundo más frecuente en cuatro diccionarios y es el más frecuente en uno de ellos; Adj Adj N que es el tercero más frecuente en todos los diccionarios del corpus; N Adj N que es el cuarto más frecuente en tres diccionarios y el quinto en dos de ellos y N N N N que es el quinto más frecuente en dos diccionarios, el cuarto más frecuente en dos diccionarios y el noveno en uno de los diccionarios. Puede decirse que salvo este último patrón descrito, estos cinco patrones son los más frecuentes en todo el corpus y casi conservan el mismo orden. Así, son las estructuras más lexicalizadas y estables de todo el corpus independientemente del área temática y el tamaño del diccionario.

Es importante destacar otras estructuras del corpus por su frecuencia de aparición en los diferentes diccionarios. Entre ellas, pueden destacarse los patrones Adj N N N, Adj Adj N N, N PPi N y N Adj N N presente dentro de los diccionarios pero con frecuencias un poco disímiles.

En general, no existe una tendencia entre los diccionarios a tener patrones exclusivos. Dentro de los primeros 20 patrones de cada diccionario hay dos diccionarios con dos patrones que no están en los 20 primeros de los otros tres diccionarios y un diccionario tiene tres patrones que no están en los otros cuatro. La excepción a esto es el diccionario IFCC que tiene nueve patrones que no aparecen en los otros cuatro diccionarios pues obedecen, en este caso, a aspectos relacionados con el área temática, el laboratorio clínico. Estos patrones tienen la categoría Num (número) en cinco de los cuatro patrones debido a que muchos términos son nomenclaturas.

En cuanto a la distribución de los primeros 20 patrones por número de tokens puede observarse que, salvo en el diccionario IFCC, los patrones más frecuentes son los de tres tokens (12, 11, 14, 11 patrones), luego siguen los patrones de cuatro tokens (7, 8, 5, 9) y, por último, los de cinco tokens (1, 1, 1, 0, respectivamente en todos los casos). En el diccionario IFCC, los patrones de cuatro tokens son los más frecuentes (13 de 20), seguidos por los patrones de tres tokens (cinco patrones) y luego los de cinco tokens (dos patrones). Igualmente, esta tendencia se rompe en este diccionario debido quizá al área temática. Sin embargo, esto último es difícil de corroborar con datos cuantitativos.

Mosby				IFCC				IMF				RD				ISI			
Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%
3	N N N	960	27,02	3	N N N	283	39	3	N N N	225	29,4	3	Adj N N	1612	30,59	3	N N N	322	26,01
3	Adj N N	913	25,7	3	Adj N N	91	12,6	3	Adj N N	209	27,3	3	N N N	1566	29,72	3	Adj N N	276	22,29
3	Adj Adj N	642	18,07	3	Adj Adj N	48	6,62	3	Adj Adj N	72	9,4	3	Adj Adj N	324	6,15	3	Adj Adj N	156	12,6
3	N Adj N	165	4,64	3	N Adj N	38	5,24	4	N N N N	43	5,61	4	N N N N	228	4,33	3	N Adj N	67	5,41
4	N N N N	79	2,22	4	N N N N	31	4,28	3	N Adj N	41	5,35	3	N Adj N	212	4,02	3	PP Adj N	44	3,55
4	Adj N N N	74	2,08	4	N Num N N	23	3,17	4	Adj N N N	20	2,61	4	Adj N N N	151	2,87	3	PP N N	39	3,15
4	Adj Adj N N	65	1,83	3	N PPi N	18	2,48	3	PP N N	20	2,61	3	PP N N	108	2,05	4	Adj Adj N N	35	2,83
3	PP Adj N	56	1,58	4	N N Adj N	11	1,52	4	Adj Adj N N	12	1,57	3	Prep N N	91	1,73	4	Adj N N N	27	2,18
3	PP N N	47	1,32	4	N Adj N N	9	1,24	3	N PP N	12	1,57	3	N PP N	85	1,61	4	N N N N	27	2,18
3	N PPi N	44	1,24	4	Adj Adj N N	8	1,1	3	PP Adj N	7	0,91	3	PP Adj N	63	1,2	4	Adj Adj Adj N	15	1,21
3	N PP N	42	1,18	4	N Adj Adj N	8	1,1	3	Adv Adj N	6	0,78	3	N PPi N	62	1,18	3	Adv Adj N	15	1,21
4	Adj Adj Adj N	36	1,01	4	Num N N N	8	1,1	3	Adv PP N	6	0,78	4	Adj Adj N N	45	0,85	3	Adv PP N	15	1,21
4	N Adj N N	22	0,62	4	N N Num N	7	0,97	4	N Adj N N	5	0,65	3	N Prep N	44	0,84	4	PP Adj Adj N	14	1,13
4	N N Adj N	21	0,59	4	Adj N N N	6	0,83	4	N Prep N N	5	0,65	3	Num N N	41	0,78	4	N Adj N N	12	0,97
4	Adj N Adj N	20	0,56	5	N N Num N N	6	0,83	4	N Adj N Num	4	0,52	5	N N N N N	38	0,72	4	Adj N Adj N	11	0,89
3	Num N N	19	0,53	4	Adj N Adj N	5	0,69	3	N PPi N	4	0,52	3	N Adv N	34	0,65	4	PP Adj N N	11	0,89
3	Adv Adj N	18	0,51	4	N N PPi N	5	0,69	4	Adv PP N N	3	0,39	3	Adv Adj N	31	0,59	4	PP N N N	11	0,89
5	Adj N N N N	17	0,48	4	N PP N N	5	0,69	5	N N N N N	3	0,39	4	N Adj N N	29	0,55	3	N PP N	10	0,81
3	PPi N N	16	0,45	5	N Num N N N	4	0,55	3	PPi N N	3	0,39	3	PPi N N	28	0,53	3	N PPi N	7	0,57
3	N Num N	13	0,37	4	N Conj N N	3	0,41	4	Adj Adj Adj N	2	0,26	4	N Prep N N	21	0,4	3	PPi N N	7	0,57

Tabla 21: Los 20 patrones más frecuentes del corpus lexicográfico en inglés.

4.4.4 Frecuencia de los patrones por longitud en diccionarios en inglés

A continuación, se presentan los patrones más comunes distribuidos por longitud de mayor a menor (+6 a 3 tokens).

Puede observarse en la tabla 22 que existe una gran variabilidad sintáctica entre los patrones de +6 tokens (7 patrones con una sola ocurrencia). Puede decirse que el patrón más frecuente es N N N N N N que aparece en 2 diccionarios (con 10 ocurrencias) y luego el patrón Adj N N N N N que aparece en 3 diccionarios (con 7 ocurrencias). Los resultados muestran que la extensión está relacionada directamente con una alta variabilidad sintáctica, como se ha planteado en varios apartados de esta tesis.

En los patrones de 5 tokens que aparecen en la tabla 23, puede observarse que aún existe una gran variabilidad sintáctica pero hay estructuras más frecuentes que aparecen en varios diccionarios del corpus lexicográfico de contraste. Así el patrón N N N N N aparece en todos los diccionarios y con una frecuencia relativamente alta, salvo en uno de los diccionarios. Luego, los patrones Adv Adj Adj N N y Adj Adj N N N aparecen en tres diccionarios y los patrones Adj N N N N, Adj N Adj N N, Adj Adj Adj Adj N y N Adj N N N en dos de los cinco diccionarios. En general, más de la mitad de los patrones de cada diccionario, salvo en el diccionario IMF, no aparecen en los otros diccionarios y son exclusivas de cada uno. En el caso de los diccionarios IFCC e ISI esta cifra es aún más alta (87% y 70%, respectivamente). En el caso del diccionario IMF, sólo un patrón no aparece en los otros, pero aún debe tenerse en cuenta que este diccionario sólo tiene cuatro patrones de cinco tokens contra los 10 patrones de media en el resto de diccionarios.

	Mosby			IFCC			IMF	Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Patrón	Frec.	%	Patrón	Frec.	%
6	Adj N N Adj N N	3	0,1	N N N N N N	3	0,4	no	N N N N N N	7	0,0013	Adj N Adj Adj N N	1	0,0008
6	Adj N N N N N	3	0,1	Adv Adj N N N N	2	0,3		Adj N N N N N	3	0,0006	Adj N N N N N	1	0,0008
6	Adj N Adj N N N	2	0,1	N N N Num N N	2	0,3		N N N N N Num N	2	0,0004	Adv Adj Adj N N N	1	0,0008

Tabla 22: Los patrones más frecuentes de más de 6 tokens del corpus lexicográfico en inglés.

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
5	Adj N N N N	17	0,48	N N Num N N	6	0,83	N N N N N	3	0,39	N N N N N	38	0,72	Adj Adj Adj N N	5	0,004
5	Adj Adj N N N	11	0,31	N Num N N N	4	0,55	Adj Adj N N N	2	0,26	Adj N N N N	11	0,0021	N N N N N	4	0,0032
5	N N N N N	10	0,28	N N N Num N	3	0,41	Adj N Adj Adj N	2	0,26	Adj Adj N N N	8	0,0015	PP Adj Adj N N	3	0,0024
5	Adj N Adj N N	9	0,25	N Num N Num N	3	0,41	Adv Adj Adj N N	2	0,26	Adj N Adj N N	7	0,0013	Adv Adj Adj N N	2	0,0016
5	Adj Adj Adj Adj N	4	0,11	Num N Num N N	3	0,41				N Adj N N N	3	0,0006	Adv Adv Adj Adj N	2	0,0016
5	PP N N N N	4	0,11	N Adj N N N	2	0,28				Num N Adj N N	3	0,0006	Adv PP Adj N N	2	0,0016
5	Adj N N Adj N	2	0,06	N Adj N Num N	2	0,28				Adj Adj Adj Adj N	2	0,0004	N Adj Adj N N	2	0,0016
5	Adj N PPi Adj N	2	0,06	N N N N N	2	0,28				Adj N PP N N	2	0,0004	N N Adj Adj N	2	0,0016
5	N PP N N N	2	0,06	N Prep N N Num	2	0,28				Adj N Prep N N	2	0,0004	PP Adj N N N	2	0,0016
5	N PPi N N N	2	0,06	Num Num Adj N N	2	0,28				Adv Adj Adj N N	2	0,0004			
5										Adv PPi N N N	2	0,0004			
5										N Adj N Adj N	2	0,0004			
5										Num N N N N	2	0,0004			

Tabla 23: Los patrones más frecuentes de 5 tokens del corpus lexicográfico en inglés.

Al igual que en los patrones de cinco tokens, los patrones de cuatro tokens tienen también variabilidad sintáctica, pero menor medida, como se ilustra en la tabla 24. De hecho, hay cuatro patrones que aparecen en todos los diccionarios y son en promedio los de mayor frecuencia entre los de cuatro tokens: N N N N, Adj N N N, Adj Adj N N y N Adj N N. Hay dos patrones que aparecen en cuatro diccionarios: N PP N N y Adj Adj Adj N.

La cantidad de patrones exclusivos es relativamente baja o nula en cuatro de los cinco diccionarios. Sólo en el diccionario Routledge, existen cuatro patrones que no aparecen en los otros cuatro diccionarios (un 20% de los 19 patrones).

En la tabla 25, se presentan los patrones de tres tokens en el corpus lexicográfico de contraste. Existen cuatro patrones que aparecen en todos los diccionarios y casi en el mismo orden de frecuencia; y, como se ha dicho antes, son a su vez, los más frecuentes en todo el corpus lexicográfico de contraste: N N N, Adj N N, Adj Adj N y N Adj N. Además, cinco patrones aparecen en cuatro diccionarios y presentan una frecuencia igualmente alta: N PPi N, Adv Adj N, N PP N y PP Adj N. Hay otros patrones de alta frecuencia en los dos diccionarios de mayor tamaño, el Routledge y el Mosby: PPi N N, Adv, PP N, PPi Adv N, entre otros. En cuanto a la exclusividad de patrones en un diccionario, puede decirse que tres diccionarios no presentan ninguna exclusividad y que los otros dos diccionarios de mayor tamaño presentan alguna exclusividad importante de patrones. En el caso del diccionario Routledge, 14 de los 27 patrones no aparecen en los otros diccionarios (51,85%) y en el diccionario Mosby hay cinco patrones exclusivos (24, 41%).

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
4	N N N N	79	2,22	N N N N	31	4,28	N N N N	43	5,61	N N N N	228	4,33	Adj Adj N N	35	2,83
4	Adj N N N	74	2,08	N Num N N	23	3,17	Adj N N N	20	2,61	Adj N N N	151	2,87	Adj N N N	27	2,18
4	Adj Adj N N	65	1,83	N N Adj N	11	1,52	Adj Adj N N	12	1,57	Adj Adj N N	45	0,85	N N N N	27	2,18
4	Adj Adj Adj N	36	1,01	N Adj N N	9	1,24	N Adj N N	5	0,65	N Adj N N	29	0,55	Adj Adj Adj N	15	1,21
4	N Adj N N	22	0,62	Adj Adj N N	8	1,1	N Prep N N	5	0,65	N Prep N N	21	0,4	PP Adj Adj N	14	1,13
4	N N Adj N	21	0,59	N Adj Adj N	8	1,1				Adj N Adj N	19	0,0036	N Adj N N	12	0,97
4	Adj N Adj N	20	0,56	Num N N N	8	1,1				N N Adj N	15	0,0028	Adj N Adj N	11	0,89
4	N Prep N N	10	0,28	N N Num N	7	0,97				N PP N N	15	0,0028	PP Adj N N	11	0,89
4	N Adj Adj N	9	0,25	Adj N N N	6	0,83				PP N N N	14	0,0027	PP N N N	11	0,89
4	PP Adj N N	8	0,23	Adj N Adj N	5	0,69				Adj Adj Adj N	13	0,0025	Adj N PP N	6	0
4	Adj N PPi N	7	0,2	N N PPi N	5	0,69				Prep N N N	11	0,0021	Adv Adj N N	6	0
4	N N PPi N	7	0,2	N PP N N	5	0,69				Num N N N	8	0,0015	N Adj Adj N	5	0
4	N PP Adj N	6	0,17							Adj N PP N	6	0,0011	N PP N N	5	0
4	N PPi N N	6	0,17							Adj N Prep N	6	0,0011			
4	PP N N N	6	0,17							N PP Adj N	6	0,0011			
4	N PP N N	5	0,14							PP Adj N N	6	0,0011			
4										Adj N PPi N	5	0,0009			
4										Adv Adj N N	5	0,0009			
4										Num Num N N	5	0,0009			

Tabla 24: Los patrones más frecuentes de 4 tokens del corpus lexicográfico en inglés.

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
3	N N N	960	27,02	N N N	283	39	N N N	225	29,37	Adj N N	1612	30,59	N N N	322	26
3	Adj N N	913	25,7	Adj N N	91	12,6	Adj N N	209	27,28	N N N	1566	29,72	Adj N N	276	22,3
3	Adj Adj N	642	18,07	Adj Adj N	48	6,62	Adj Adj N	72	9,4	Adj Adj N	324	6,15	Adj Adj N	156	12,6
3	N Adj N	165	4,64	N Adj N	38	5,24	N Adj N	41	5,35	N Adj N	212	4,02	N Adj N	67	5,41
3	PP Adj N	56	1,58	N PPi N	18	2,48	PP N N	20	2,61	PP N N	108	2,05	PP Adj N	44	3,55
3	PP N N	47	1,32				N PP N	12	1,57	Prep N N	91	1,73	PP N N	39	3,15
3	N PPi N	44	1,24				PP Adj N	7	0,91	N PP N	85	1,61	Adv Adj N	15	1,21
3	N PP N	42	1,18				Adv Adj N	6	0,78	PP Adj N	63	1,2	Adv PP N	15	1,21
3	Num N N	19	0,53				Adv PP N	6	0,78	N PPi N	62	1,18	N PP N	10	0,81
3	Adv Adj N	18	0,51							N Prep N	44	0,84	N PPi N	7	0,57
3	PPi N N	16	0,45							Num N N	41	0,78	PPi N N	7	0,57
3	N Num N	13	0,37							N Adv N	34	0,65			
3	Num Adj N	13	0,37							Adv Adj N	31	0,59			
3	Prep N N	7	0,2							PPi N N	28	0,53			
3	N N Adj	6	0,17							Adv PP N	20	0,0038			
3	Adj N PP	5	0,14							PP Prep N	15	0,0028			
3	Adj PPi N	5	0,14							PPi Adv N	10	0,0019			
3										Adv N N	9	0,0017			
3										Adj Adv N	8	0,0015			
3										PP Adv N	7	0,0013			
3										PPi Prep N	7	0,0013			
3										Num Adj N	6	0,0011			
3										Adj Prep N	5	0,0009			
3										N Num N	5	0,0009			
3										Prep Adj N	5	0,0009			
3										V Adj N	5	0,0009			
3										V Adv N	5	0,0009			

Tabla 25: Los patrones más frecuentes de 3 tokens del corpus lexicográfico en inglés.

4.5 Contraste de resultados entre el corpus de análisis y el corpus lexicográfico en inglés

A continuación, se comparan los resultados obtenidos en el corpus de análisis en inglés y el corpus lexicográfico de contraste en cuanto a la distribución por longitud, la categoría léxica predominante en la premodificación, la frecuencia de patrones por aparición y la frecuencia de los patrones de acuerdo con la longitud.

4.5 1 Distribución de acuerdo con la longitud

En la tabla 26 la comparación de los datos de ambos corpus, permite establecer que efectivamente, las tendencias presentadas en ambos casos, y con porcentajes tan similares, muestran que las estructuras más estables están directamente relacionadas con una menor extensión. Debe tenerse en cuenta que los diccionarios presentados en esta tesis no se confeccionaron con metodología de corpus como suele hacerse actualmente con los diccionarios generales en inglés y algunos técnicos. Por tanto, puede decirse que los diccionarios pueden ser un reflejo de la lengua en este sentido y no distan significativamente de los datos del corpus. Desde un punto de vista sintáctico, la intuición lingüística del lexicógrafo se ve corroborada en los datos. Es decir, la selección de unidades con estas estructuras no es caprichosa sino que responde a las tendencias lingüísticas del hablante.

	Corpus		Diccionarios	
N.º tokens	Frecuencia	Porcentaje	Frecuencia	Porcentaje
3	909	86,16	9.561	82,77
4	135	12,8	1.632	14,12
5	11	1,04	286	2,47
6	0	0	59	0,51
7	0	0	13	0,11
Total	1.055	100	11.551	100

Tabla 26: Frecuencia por número de tokens entre el corpus de análisis y el corpus lexicográfico en inglés.

Como ya se dijo en §4.3.1, las estructuras del corpus que tienen esta tendencia en la longitud (de 3 a 4 tokens) tenderán a ser más estables y son las que potencialmente pueden convertirse en términos en los diccionarios, como se infiere de los resultados del corpus de diccionarios. Desde un punto de vista lexicográfico, estos datos son muy relevantes para la confección de diccionarios ya que son un parámetro más en la selección de unidades candidatas a término.

4.5.2 Categoría léxica predominante y aspectos morfológicos

Al igual que en el corpus de análisis, el corpus lexicográfico presenta casi el mismo porcentaje de sustantivos en la premodificación de los sintagmas como puede verse en la tabla 27. Además, casi doblan en porcentaje a los adjetivos, lo que corrobora las observaciones hechas antes (Quiroz 2005). En este sentido, puede decirse que en los diccionarios también se reflejan las tendencias léxicas y gramaticales del discurso especializado, y como se dijo antes, dicha tendencia en el uso de la categoría léxica no es dependiente del área temática concreta sino quizá del discurso especializado en general.

	Corpus	Diccionarios
Cat. léxica	Porcentaje	Porcentaje
N (sin núcleos)	45,95	44,15
Adj	32,43	25,02
PP	13,51	7,47
PPi	N.D[50]	4,59
Adv	8,11	6,95
otras	0	11,82

Tabla 27: Comparación de la categoría léxica predominante entre el corpus de análisis y el corpus lexicográfico en inglés.

Los sustantivos representan casi la mitad de las unidades léxicas que aparecen en la premodificación de ambos corpus. Morfológicamente, la mayoría de sustantivos son derivados deverbales terminados en *-ation* con 88/62 sustantivos en el corpus de análisis y 1.021/285 en el corpus lexicográfico en los núcleos y la premodificación, respectivamente. En segundo lugar, aparecen los sustantivos derivados deadjetivales terminados en *-ity* con 20/10 y 179/86, *-er* con 20/15 y 250/102 sustantivos en ambos corpus, respectivamente. Como puede verse estos sufijos son deverbales y como se observa en los siguientes casos de los ejemplos 31 y 32 y deadjetivales como en los casos de 33.

31. association, organization, corporation, concentration, operation, transformation, hybridization, insemination, inhalation, donation, sterilization, amalgamation, estimation

32. manager, container, officer, worker, practitioner, printer, trader, carrier, converter, multiplier, analyzer, computer, counter, dealer

33. facility, security, capacity, inequality, liability, personality, hypersensitivity, activity, deformity, convertibility

[50] Por razones de etiquetaje del corpus no fue posible separar los participios de presente y pasado como en el corpus lexicográfico de contraste. Obsérvese que los porcentajes son similares en ambos casos si se suman los del corpus lexicográfico.

Estos datos corroboran el mismo orden de productividad observado por Biber *et al* (1999: 322-323) para el discurso académico y en el conjunto de la lengua inglesa, el discurso académico tiene las frecuencias más altas de este tipo de sufijos derivativos. Como bien lo plantea Biber *et al* (1999: 325) la presencia de estos es básica en este tipo de discurso ya que se hace frecuente referencia a conceptos abstractos y en donde se referencian las acciones y procesos en términos generales en vez de estar relacionados a un tiempo y un lugar determinados. En este sentido, es conveniente el uso de nominalizaciones para comprimir en un sintagma nominal el contenido de una cláusula. Así, la nominalización es el recurso más eficiente dentro del discurso especializado (Iturrioz 2000; Gallegos 2000, 2003) y esto se observa en la cantidad de sustantivos no sólo en los núcleos sino en la premodificación. Su uso se justifica pragmáticamente por los objetivos que se persiguen en la ciencia.

A continuación, aparecen los adjetivos con un tercio de las unidades léxicas en el corpus de análisis y un cuarto en el corpus lexicográfico. Esta diferencia se explica debido a que en el corpus lexicográfico hay un 11% de patrones con otras categorías léxicas como verbo, preposición, etc.

Luego, siguen los participios de presente (PPi) y de pasado (PP) con aproximadamente un 13% en ambos corpus. En el corpus de análisis no aparecen separadas categorías debido a que no estaban claramente diferenciadas en el etiquetaje. Como ya se explicó en §4.3, los participios de pasado son más frecuentes que los participios de presente (7,47 vs. 4,59) en el corpus lexicográfico. Los PP son en su mayoría deverbales ya que resultan de una oración relativa pasiva como ocurre en el ejemplo 34.

34. bidirectionally determined restriction sites (restriction sites which are bidirectionally determined)

Los participios de presente, cuya función está más cerca de los sustantivos que de los adjetivos y en los patrones más frecuentes, se encuentran

inmediatamente al lado del núcleo y al lado de otro sustantivo, como ocurre en los ejemplos 35 y 36.

35. osteoclast activating factor (N PPi N)
36. ionizing radiation injury (PPi N N)

En el caso de los adverbios, en su mayoría se derivan de adjetivos y estos de sustantivos, como puede verse en los ejemplos 37 y 38.

37. environmentally-friendly product (environmentally > environmental > environment)
38. genetically determined immune response (genetically > genetic > genetics)

Morfológicamente, la mayoría de adverbios terminan en *-ly* como los casos del ejemplo 39 y otros son locuciones adverbiales que provienen del latín como en los ejemplos de 40.

39. highly, biologically, alternatively, bidirectionally, genetically, maternally, oxidatively, phylogenetically, physically, anatomically, biomedically, chemically, covalently, exponentially, homogeneously, morphologically, pharmaceutically, phenotypically

40. in situ, in vitro, in vivo, ex vivo

Desde un punto de vista sintáctico, estos adverbios y los participios de pasado o adjetivos que generalmente los acompañan son también el resultado de una oración relativa pasiva como en el ejemplo 41.

41. genetically determined immune response (immune response which is genetically determined)

Esta tendencia en ambos corpus puede explicarse desde varios puntos de vista.

La tendencia nominalizadora del discurso especializado como una forma de mostrar objetividad se ve reflejada en la gran cantidad de sustantivos en la premodificación y son los sintagmas los que llevan esta carga como lo propone Vivanco (1996: 752).

> Con el fin de incorporar en un texto científico la objetividad, impersonalidad y generalidad, comúnmente relacionada con el uso de largos y complejos GNs, a nombres abstractos y a nominalizaciones, el científico elige como medio de transmisión, el lenguaje nominal.

Esto se evidencia en la cantidad de sustantivos terminados en *–tion*, 62 de los 1.296 sustantivos en posición premodificadora en el corpus de medicina. Morfológicamente, es el sufijo más productivo en la muestra. Téngase en cuenta que en la tabla 27 sólo se ha descrito la premodificación ya que el núcleo siempre será un sustantivo.

La lexicalización de muchos participios de presente como verdaderos sustantivos se observa en el hecho de ser entradas o subentradas, como se observa en los ejemplos de 42, los cuales se encuentran en algunos diccionarios de referencia[51].

42. binding, mapping, sensing, processing, screening, imprinting, etc.

4.5.3 Frecuencia de los patrones por aparición

En este apartado, se comparan los 20 patrones más frecuentes de ambos corpus. Puede observarse en la tabla 28 que en ambos corpus el patrón más frecuente es el patrón N N N con casi el mismo porcentaje de ocurrencias (29% aprox.). En segundo lugar, el patrón Adj N N aparece con porcentajes absolutos

[51] Webster's New World Dictionary & Thesaurus en CD-ROM, 1998.

también muy similares entre ambos corpus (24,08% y 26,85%). Luego, aparece el patrón Adj Adj N con porcentajes casi idénticos en ambos corpus (10,71% y 10,75%) y en cuarto lugar aparece el patrón N Adj N con porcentajes similares (5,88% y 4,53%). Posteriormente, el orden en ambos corpus comienza a variar pero en los primeros 10 patrones hay 9 patrones iguales y 1 diferente. En los segundos 10 patrones hay 3 patrones iguales y el resto no coinciden independientemente del orden, pero 2 de ellos coinciden con los 2 de la primera franja.

Tokens	Patrón corpus	Porcentaje	Tokens	Patrón Dic.	Porcentaje
3	N N N	30,05	3	N N N	29,05
3	Adj N N	24,08	3	Adj N N	26,85
3	Adj Adj N	10,71	3	Adj Adj N	10,75
3	N Adj N	5,88	3	N Adj N	4,53
3	PP N N	4,83	4	N N N N	3,54
4	Adj N N N	3,13	4	Adj N N N	2,41
3	PP Adj N	2,94	3	PP N N	1,88
4	N N N N	2,65	3	PP Adj N	1,49
3	Adv Adj N	2,37	4	Adj Adj N N	1,43
3	N PP N	2,18	3	N PP N	1,32
4	Adj Adj N N	1,52	3	N PPi N	1,18
3	Adv PP N	1,8	3	Prep N N	0,88
3	Adj PP N	1,23	4	N Adj N N	0,62
4	N Adj N N	0,95	3	Adv Adj N	0,61
4	Adj N Adj N	0,85	4	Adj Adj Adj N	0,59
4	PP N N N	0,57	3	Num N N	0,52
4	PP Adj N N	0,47	5	N N N N N	0,49
4	Adj Adj Adj N	0,38	3	PPi N N	0,49
4	Adv Adj N N	0,38	4	Adj N Adj N	0,48
4	Adv PP N N	0,38	4	N N Adj N	0,44

Tabla 28: Comparación de los primeros 20 patrones del corpus y el corpus lexicográfico en inglés.

Los 33 patrones del corpus de análisis aparecen en todos los diccionarios, pero no al contrario debido a que se extrajo una muestra de éstos al ser un corpus de análisis. Es decir, el corpus lexicográfico se usó para corroborar que los datos del corpus de análisis no eran arbitrarios y correspondían con los de la lengua y no para ver si los patrones recogidos por los lexicógrafos eran realmente usados por los especialistas en discurso.

Obsérvese que los 4 primeros patrones que coinciden plenamente en orden y casi en frecuencia de aparición agrupan el 70% aproximadamente de todas las ocurrencias de ambos corpus. Esto muestra, por un lado, la alta productividad respecto del resto de patrones, su alta probabilidad de ocurrir en el discurso y en los diccionarios y, por otro lado, su grado de estabilización como estructuras en el discurso especializado es evidente, lo que puede traducirse en varios aspectos: 1) las unidades léxicas de un texto que tengan estas cuatro estructuras tenderán a ser firmes candidatas a término como puede verse en los ejemplos extraídos del corpus; 2) las regularidades que se observen en el corpus paralelo de §8 sobre estos cuatro patrones pueden ser igualmente regulares, lo que sugerirá un comportamiento determinado de ellas en español; 3) los sistemas de extracción lingüísticos o híbridos pueden dar un peso relevante a estos cuatro patrones para mejorar los resultados de una extracción terminológica; 4) las consecuencias didácticas para la traducción y la enseñanza de los LSP son importantes en cuanto a que estos resultados pueden ayudar en la selección de casos frecuentes para traducir o analizar cuando se enseñen aspectos lingüísticos o problemas frecuentes del discurso académico.

4.5.4 Frecuencia de los patrones por longitud

En cuanto a su longitud, puede observarse que en la franja de los 10 primeros patrones predominan los patrones de 3 tokens de extensión en ambos corpus. En el corpus de análisis hay 8 patrones de 3 tokens contra 2 de 4 tokens, y en el corpus lexicográfico existe un patrón más de 4 tokens.

En la segunda franja de 10 patrones, existe un predominio de los patrones de 4 tokens de extensión en el corpus de análisis y no existe un predominio claro en el corpus lexicográfico. Al contrario, en la primera franja, en el corpus de análisis hay 8 patrones de 4 tokens y 2 de 3 tokens. En el corpus

lexicográfico hay 4 patrones de 4 tokens, 5 de 3 tokens y 1 de 5 tokens (N N N N N).

En general, puede decirse que predominan los patrones de 3 tokens y luego siguen los patrones de 4 tokens. Esto, muestra una vez más, que la extensión está ligada a la frecuencia, como se explicó en §4.3.3.

De la comparación de ambos corpus, puede desprenderse que salvo en unos pocos casos, los análisis realizados coinciden plenamente no sólo en el orden de preferencia de los corpus de las categorías léxicas como en patrones más regulares, sino en la frecuencia de aparición de estas categorías y estructuras.

4.6 Recapitulación

En este capítulo, se han presentado los resultados del análisis formal del corpus de análisis en inglés y el contraste con el corpus lexicográfico.

1. En cuanto a la longitud de los sintagmas, el corpus de análisis en inglés, los patrones de 3 tokens predominan ampliamente sobre la demás longitud con un 86,16% de todas las ocurrencias, seguidos de los patrones de 4 tokens con 12,8%.

2. En cuanto a la categoría gramatical predominante en la premodificación, se confirman las tendencias que se han obtenido en el estudio piloto en el uso del sustantivo como premodificador con un 45,95%, seguido por el adjetivo con un 32,43%. Es importante resaltar la presencia de otras categorías léxicas como los participios con un 13,51% y los adverbios con 8,11%. Este alto uso del sustantivo en la premodificación refuerza el carácter nominalizador e impersonal del discurso científico- técnico.

3. En cuanto a los patrones más frecuentes, los patrones más frecuentes son N N N con un 30,05%, Adj N N con un 24,08%, Adj Adj N con un 10,71% y el patrón N Adj N con 5,88%. Estos cuatro patrones representan el 70,72% de todas las ocurrencias del corpus y por tanto, presentan menos variación sintáctica. Entre los patrones de 4 tokens cabe destacar los patrones Adj N N N con 3,13 y N N N N con 2,65%. Estos patrones presentan más variación y representan a muchas menos ocurrencias.

4. De acuerdo con la dependencia sintáctica, la relación de dependencia [C [[B A]] es la más frecuente en todo el corpus en inglés con más del 61% de todas las ocurrencias (142) del corpus de análisis, seguida de la relación de dependencia [[C B] A] con un 28,89% de todas las ocurrencias (67). Por último,

la dependencia [[D C] [B A]] representa el 5,17% de todas las ocurrencias (12) para patrones de 4 tokens. Los patrones que presentan una única relación de dependencias son: Adj Adj N, N Adj N, N PP N, PP Adj N, PP N N, Adv Adj N, Adj N N N y N N N N. De estos, los patrones Adj Adj N, N Adj N, N PP N, PP Adj N, PP N N tienen la misma relación de dependencia sintáctica [C [[B A]]. El único patrón que tiene la dependencia sintáctica [[C B] A] es Adv Adj N. En los dos patrones de 4 tokens, Adj N N N y N N N N, la dependencia que predomina es [[D C] [B A]]. Los patrones que tienen dos relaciones de dependencia sintáctica son: Adj N N y N N N. En el patrón Adj N N, la relación de dependencia [C [[B A]] representa al 64,78% y [[C B] A] al 35,21%. En el caso del patrón N N N, la dependencia [[C B] A] representa al 87,93% y la dependencia [C [[B A]] al 12,07%.

5. El contraste con el corpus lexicográfico corrobora los resultados obtenidos en el corpus de análisis en cuanto a la longitud y frecuencia de los sintagmas, predominio de patrones y categoría léxica en la premodificación. Así, se puede afirmar que los análisis hechos se pueden extrapolar a otras áreas del conocimiento y que no son exclusivos de las ciencias “duras”.

Así, se ha demostrado que la existencia de los SNEE es una característica de la lengua que puede presentarse con mayor frecuencia en el discurso especializado y que, además, pueden describirse, clasificarse, explicarse y predecirse desde la gramática de una lengua como todos los fenómenos lingüísticos de los discursos de los ámbitos de especialidad, como lo plantea la teoría comunicativa de la terminología.

5. Análisis formal de los patrones en español

5.1 Introducción

Los sintagmas nominales extensos han sido poco estudiados desde la gramática o la lingüística española. Este mismo hecho no ocurre en otras lenguas como el inglés y el francés, como se muestra en el §2. Cómo bien lo manifiesta Montero (1995: 45):

> "En lo que respecta a la lengua española, el tema aparece, desde luego, en las gramáticas u obras generales, que dedican algún capítulo al nombre y a las palabras que pueden modificar su significado. En ellas se estudian los componentes del sintagma nominal por separado y al considerarlos como grupo, se centran fundamentalmente en su tipología, género, número y problemas de concordancia que pueden plantear."

Incluso los estudios que se han hecho desde otras disciplinas han sido más bien incipientes o tangenciales. En terminología, este fenómeno ha despertado poco interés y sólo en los últimos años algunos autores como Montero (1995) Cardero (2000, 2004), León (2003), Oster (2005), Cortés (2004) y Quiroz (2005) han rescatado su importancia dentro de la descripción lingüística y funcional de los sintagmas nominales extensos que son términos y de las terminologías de diferentes ámbitos y desde diferentes lenguas hacia el español. Otros estudios terminológicos (Estopà 1999 y Vivaldi 2004) han reconocido su existencia, pero no las han estudiado por no ser su objeto de estudio. Desde un punto de vista fraseológico, los estudios como el de Belvilacqua (2004) han hecho importantes aportes a las unidades sintagmáticas de carácter eventivo. También son relevantes para este trabajo, los trabajos llevados a cabo en el marco de la TCT en cuanto a la descripción de categorías léxicas que son parte del sintagma como los trabajos sobre adjetivos de Folguerà (2002), de participios de Salazar (2006), de siglas (Giraldo 2005), de recuperación de sintagmas extensos (Quiroz 2005), entre otros.

En traducción, no hay muchos trabajos que estudien el fenómeno descriptivamente. Ningún manual de traducción al español trata la traducción de este tipo de unidades adecuadamente, salvo los trabajos de Gallardo (1997), Linder (2002) que intentan sistematizar este tipo de sintagmas y proponen algunas alternativas aunque sin mucho acierto como en el caso de este último. Su tratamiento no responde a una observación sistemática de datos y cómo había sido tradicional hasta hace unos años, los trabajos se dedicaban a dar instrucciones con base en la intuición del hablante u observaciones casuísticas. En cambio, son muchos los autores que reconocen el fenómeno como un problema o una característica del discurso científico-técnico, pero sin proponer una solución de ellos en español (Vázquez-Ayora 1977; López y Minett 1997; Abril y Ortiz 1998; Cartagena 1998; Quiroz y Muñoz 1997; entre otros).

Sin embargo, las descripciones hechas por los gramáticos clásicos y otros autores en español (Bosque 1999; Alcina-Blecua 1975; Lacuesta y Bustos 1999: 4505; Rainer 1999: 4595; Varela y Martín 1999: 4993; Demonte 1999: 128) en cuanto a la descripción de sintagmas nominales de poca extensión y al papel que juegan las diferentes categorías léxicas dentro del sintagma, o fuera de él, son de especial relevancia para describir los sintagmas nominales extensos en español.

En español un sintagma nominal extenso especializado es una proyección de un sustantivo núcleo que tiene más de dos modificadores de categoría léxica abierta, bien sean sintagmas preposicionales como en el ejemplo 1, adjetivos como en el ejemplo 2, adjetivos y participios como en el ejemplo 3, y adjetivos modificados adverbialmente como en el ejemplo 4.

1. electroforesis **en gel de agarosa**
2. poliquistosis **renal autosómica recesiva**
3. células **alveolares descamadas**
4. loci **altamente polimórficos**

Puesto que es posible encontrar algunos patrones con adjetivos antepuestos o pospuestos como en los ejemplos 5 y 6, no es conveniente referirse a posmodificación, si bien la mayor parte de los patrones de este estudio están posmodificados. Por tanto, se empleará la denominación "modificadores del sustantivo (o nombre)" para hacer referencia a todo lo que no sea el núcleo.

5. **alto** grado de polimorfismo
6. hepatitis **vírica crónica**

5.2 Criterios y selección del corpus de análisis en español

Al igual que en inglés, se extrajo una muestra con un 3% de error del corpus de análisis del español de 1.081 ocurrencias, para el corpus de análisis de los patrones de acuerdo con lo explicado en el capítulo de la metodología (§3.7). Cada sintagma de la muestra se seleccionó manualmente y se distribuyó proporcionalmente de acuerdo con la frecuencia de aparición de cada patrón. Por ejemplo, el patrón más frecuente fue N Prep N Adj con más de 7.963 ocurrencias en el corpus total y, por tanto, le correspondía una muestra de 225 sintagmas de los 1.081 sintagmas totales. En cambio, uno de los 3 patrones con menos frecuencia es el N N Prep N Prep N Adj (5) y le correspondía 1 sintagma de muestra.

Asimismo, dentro de cada patrón se seleccionaron los sintagmas de mayor a menor frecuencia hasta completar la muestra que correspondía a cada patrón sintáctico. Por ejemplo, en el patrón N Prep N Adj, uno de los sintagmas más frecuentes es *gen de la fibrosis quística* con 7 ocurrencias y, uno de los menos frecuentes de la muestra es *virus de la leucemia bovina* con 1 ocurrencia.

A continuación, se llevó a cabo una estadística descriptiva de la muestra en cuanto a la distribución por longitud de los sintagmas en el corpus, categoría léxica predominante, patrones más frecuentes en la muestra, patrones más frecuentes por extensión, patrones más frecuentes por categoría léxica.

Posteriormente, se analizó el corpus lexicográfico de contraste bajo los mismos parámetros empleado para el corpus de análisis. Sin embargo, en cada análisis cuantitativo se comparan los cinco diccionarios entre sí.

A partir de estos dos análisis, se comparan los datos obtenidos en la muestra del corpus de referencia con los del corpus lexicográfico de contraste y se analizan a la luz de los resultados obtenidos por otros autores en algunos de los análisis llevados a cabo aquí.

Finalmente, se comparan los patrones del corpus de análisis y los resultados del corpus lexicográfico con los datos obtenidos del corpus CREA de la Real Academia Española[52].

Así, pues, en este capítulo se pretende demostrar que los sintagmas nominales extensos no son un fenómeno raro de la lengua española, hecho que se ve reflejado no sólo en el corpus de análisis y en el corpus lexicográfico de contraste sino también en el corpus CREA de la RAE. Por tanto, las tendencias que presenten los patrones se pueden describir, explicar y predecir desde la gramática general del español.

[52] Datos obtenidos del Banco de datos (CREA), *Corpus de referencia del español actual* por gentileza de la Real Academia Española, [Consulta recibida el 20.02.2007].

5.3 Resultados

Con la muestra antes descrita, se tabularon todos los datos en el programa de estadística Statgraphics Plus 5.1 para obtener los resultados que se describen a continuación.

5.3.1 Longitud y frecuencia de los SN en español

En el corpus de análisis se extrajeron unidades desde 3 tokens (uno como núcleo) hasta 6 tokens, como se ilustra en los ejemplos 7 y 8.

7. virus de la inmunodeficiencia humana (N Prep N Adj)
8. actividad de la enzima responsable de la síntesis de óxido nítrico (N Prep N Adj Prep N Prep N Adj)

Al igual que en inglés, se encontraron unidades aún más extensas en otros autores (Guzmán 2002; Oster 2003; Cortés 2005) y cuyos datos se extrajeron manualmente.

Como puede verse en la tabla 1, los patrones de 3 tokens (dos en la premodificación) son los más frecuentes en la muestra con un 80,66% del total (872 ocurrencias). Por el contrario, los sintagmas de 4, 5 y 6 tokens tan sólo representan un 19,32% del total de sintagmas (14,74%, 2,49% y 0,09%, respectivamente).

N.º tokens	Frecuencia	Porcentaje
3	872	80,66
4	181	16,74
5	27	2,496
6	1	0,09
Total	1.081	100

Tabla 1: Frecuencia por número de tokens del corpus de análisis del español.

En el corpus general del español, también se extrajeron patrones de seis y siete tokens, pero no se han incluido por su baja frecuencia ya que el criterio de inclusión fue de más de cinco ocurrencias en la muestra.

Como puede verse en los estudios presentados en §2, el corpus de este estudio presenta una tendencia similar al estudio de Guzmán (2003), es decir, a menor extensión del patrón, mayor frecuencia de aparición. No es casualidad que la muestra sólo contenga patrones de 3, 4 y 5 tokens ya que estas estructuras son las que pueden revertir más en estabilización y posible lexicalización del sintagma, como lo plantea Cartagena (1998).

5.3.2 Categoría léxica predominante en la posmodificación

Al contrario que en inglés, la categoría léxica por excelencia para modificar al núcleo es el adjetivo en español. De hecho, una de las estructuras más estudiadas en la gramática y en terminología es el patrón N Adj y Adj N. A diferencia del inglés, en español no se establece que el adjetivo sea más frecuente como modificador del sustantivo dentro de un sintagma. Sin embargo, puede deducirse de los estudios realizados que se asume que es la categoría preferida.

POS	Frecuencia	Porcentaje
N (sin núcleos)	1.220	30,37
Adj	1.077	26,81
PP	103	2,56
Adv	13	0,32
Prep	1.184	29,47
D	419	10,43
Total	4.016	99,96

Tabla 2: Categoría léxica predominante en la posmodificación.

Como puede verse en la tabla 2, puede decirse que en el corpus de análisis predominan los sustantivos como categoría modificadora del sustantivo

núcleo con un 30,37% contra 26,81%. Incluso si se sumaran los participios de pasado como potenciales adjetivos, se obtendría un 29% aproximado. Esta tendencia muestra la aparición de más sustantivos como categoría modificadora en español, al menos en este tipo de discurso.

Esta tendencia en el uso de más sustantivos, se debe a varios factores. En primer lugar, fuera de los tipos de sintagmas SA y AS (N Adj y Adj N), la estructura que tiende a predominar en la formación de sintagmas es el uso de complementos preposicionales (Prep SN). Por tanto, las posibilidades aumentan considerablemente. En segundo lugar, dado el carácter nominalizador del discurso científico para representar objetos, eventos y procesos, el sustantivo es la categoría léxica por excelencia, para vehicular el conocimiento especializado; lugar propicio para concentrar los nudos de conocimiento.

Los adjetivos, como es de esperarse, son la segunda categoría abierta más frecuente con un 26,81%. Un aspecto importante es que hay un 2,87% de adjetivos en posición premodificadora (31 ocurrencias).

Esta frecuencia similar a la del sustantivo, muestra la importancia que tiene el adjetivo como elemento modificador del núcleo. Así, Folguerà (2002: 213-215) muestra la importancia del adjetivo dentro del discurso especializado no sólo cómo una unidad discursiva autónoma sino como parte de un sintagma. Propone que los adjetivos que no son lexemáticamente especializados, es decir, los adjetivos de dimensión o temporales (corto, largo, pequeño, etc.) dentro del sintagma pueden transmitir conocimiento especializado (más de un 50% de ellos). Además, muestra la importancia de los adjetivos como elementos lexicalizadores de un sintagma (2002: 206). Establece que, para que una estructura de tipo N Adj se lexicalice, debe cumplir tres requisitos: 1. que el núcleo sea una unidad especializada (término), 2. que la secuencia presente antonimia, y 3. que el adjetivo sea clasificador. Si uno de estos 3 parámetros no se cumple el sintagma simplemente mantendrá una cohesión colocacional (2002: 211).

Para Salazar (2006: 73) la categoría gramatical que mejor caracteriza lingüísticamente la diferencia del discurso de la economía respecto al discurso de lengua general es la adjetival. En cuanto a los participios de pasado del corpus de análisis en español de esta tesis, puede observarse que son relativamente pocos (2,56%), pero si se tiene en cuenta que es una categoría poco frecuente dentro de los sintagmas, su importancia como elemento lingüístico es fundamental en la formación de nuevos términos. Salazar (2006: 26) muestra que los participios son los adjetivos deverbales más frecuentes no sólo en un corpus de lengua general sino en un diccionario de economía.

Kornfeld & Resnik (2002: 1) plantean que a pesar de ser muy productivos, los participios han sido poco estudiados en español:

> "Estos adjetivos, que son muy productivos en español, han sido paradójicamente muy poco estudiados en la bibliografía sobre morfología del español. Ello puede deberse a su naturaleza categorial ambigua y al hecho de que la forma participial verbal, que participa de las construcciones pasivas, ha sido estudiada desde la sintaxis."

En cuanto a las preposiciones, debido a la cantidad de complementos preposicionales, no sólo aumenta la cantidad de sustantivos sino de preposiciones, que representan un 28,35%. De hecho, es la segunda categoría más frecuente después de los sustantivos en el corpus de análisis del español.

5.3.3 Frecuencia de los patrones por aparición

A diferencia del inglés, la distribución de patrones por número de tokens de la tabla 3 muestra que los patrones más frecuentes son los de 4 tokens con casi la mitad de los patrones (47,05) y una media de 6,46 sintagmas por patrón. Luego siguen los patrones de 3 tokens con casi un tercio de los patrones

(29,42%) y con una media de 51,29 sintagmas por patrón. En tercer lugar están los patrones de 5 tokens con un poco más de un quinto de los patrones (22,05%) y una media de 1,9 sintagmas por patrón. Y por último están los patrones de 6 tokens con un 1,47% de los patrones y un promedio de 1 patrón por cada sintagma.

Es importante tener en cuenta que se crearon más patrones de 4 tokens para interrogar el corpus del IULA y el CREA de la RAE (18 de 3 tokens, 32 de 4, 21 de 5 y 1 de 6 tokens). A pesar de que hay más patrones de 4 tokens que de 3, estos representan muchas menos ocurrencias (181 vs. 872). Estos datos muestran, como sucede en inglés, que la variabilidad en los patrones de superficie menos extensos es menor si se compara con la alta variabilidad de los patrones más extensos, que es de casi un patrón por cada dos sintagmas. Desde un punto de vista traductivo, esta variabilidad en los patrones más extensos dificulta la sistematización de este tipo de patrones puesto que son precisamente estos los que más problemas de traducción presentan.

N.º tokens	Patrones	Porcentaje	Frecuencia
3	17	29,42	872
4	28	47,05	181
5	14	22,05	27
6	1	1,47	1
Total	68	100	1.081

Tabla 3: Distribución de patrones por extensión en el corpus de análisis en español.

En la tabla 4 se presentan los 20 patrones más frecuentes de la muestra. Para el análisis se han separado en dos franjas de 10 patrones. Estos 20 patrones representan el 92,83% del total de la muestra con 1.009 ocurrencias sobre un total de 60 patrones y 1.081 ocurrencias. Es decir, que, en el resto de la muestra, existe una gran variabilidad en los 40 patrones restantes y 72 ocurrencias; es decir, hay casi un patrón por cada 1,8 sintagmas. En cambio, en estos 20 patrones existe una relación de 1 patrón por cada 50 sintagmas, lo que permite llevar a cabo generalizaciones más confiables, cuestión que no es muy factible con los otros 40 patrones.

De igual modo, puede verse que los primeros 10 patrones representan el 84,34% de toda la muestra con 917 sintagmas sobre los 92 sintagmas de los segundos 10 patrones (8,49%).

Entre los primeros 20, existe un predominio de los patrones de 3 tokens con 12 patrones, seguidos de los patrones de 4 tokens con 7 patrones y 1 patrón de 5 tokens. En ambas franjas predominan los patrones de 3 tokens (7 en la primera y 5 en la segunda franja).

En toda la muestra, los 3 patrones más frecuentes son N Prep N Adj, N Adj Prep N y N Prep N Prep N y, en conjunto, agrupan 636 sintagmas que representan un 58,47% del total de la muestra y, por extensión, del corpus de análisis.

Longitud	Patrón	Ejemplo	Frec.	Porc.
3	N Prep N Adj	virus de la inmunodeficiencia humana	343	31,66
3	N Adj Prep N	artrosis degenerativa de la columna	175	16,13
3	N Prep N Prep N	electroforesis en gel de agarosa	118	10,68
3	N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
3	N Adj PP	células alveolares descamadas	53	4,88
4	N Adj Prep N Adj	membrana apical de las células epiteliales	40	3,68
4	N Prep N Adj Prep N	constricción de las arterias coronarias del corazón	36	3,31
3	Adj N Prep N	alto grado de polimorfismo	30	2,76
4	N Adj Prep N Prep N	secreción excesiva de hormona de crecimiento	29	2,67
3	N Prep Adj N	sulfonilurea de alta afinidad	20	1,84
3	N PP Prep N	oligonucleótidos repetidos en tándem	14	1,29
4	N Prep N Prep N Prep N	hipocrecimiento por anomalías en genes de los gonosomas	12	1,11
4	N Prep N Prep N Adj	electroforesis en geles de campos pulsantes	11	1,01
3	N Prep N PP	hibridación con sonda marcada	10	0,93
4	N Prep N Adj Adj	inoculación con adenopatías satélites axilares	9	0,83
4	N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74
3	N N Adj	hormona somatomamotropina coriónica	8	0,74
5	N Adj Prep N Prep N Adj	síndrome dismetabólico de sobrecarga de hierro heterocigoto	7	0,64
3	N Adv Adj	loci altamente polimórficos	7	0,65
3	N Adj N	hepatitis vírica C	6	0,55

Tabla 4: Los 20 patrones más frecuentes del corpus de análisis en español.

A continuación se analizan cuantitativamente los 20 patrones más frecuentes de acuerdo con el tipo de categorías léxicas modificadoras: con sustantivo, con adjetivo, con participio, con adverbio, sin sustantivo, sin adjetivo y sin preposición.

En las tablas de este ítem no se van a incluir patrones de una única frecuencia salvo en los patrones con baja frecuencia (p. ej., participios, adverbios y preposiciones), pero se especificará en cada caso el número de patrones totales y los porcentajes y ocurrencias a las que representan. También, se ha eliminado la categoría D (determinante) para resumir la cantidad de patrones y poder hacer más generalizaciones ya que hay patrones que solo se diferencian por el uso del determinante, aunque somos conscientes de que la inserción de un determinante puede incidir en la no lexicalización de un SN como lo advierte Alvar (1993: 23):

> "En los compuestos por sinapsia, la segunda parte, el elemento determinante, carece de artículo, pues de lo contrario rompería la unidad del conjunto".

Sin embargo, puede verse que muchos de los términos que se encuentran en el corpus lexicográfico tienen determinantes, como sucede en el ejemplo 9.

9. virus de la inmunodeficiencia humana

También hay sintagmas que se encuentran en el diccionario sin determinante, pero que en otros se encuentra con él, como en el ejemplo 10:

10. secreción excesiva de hormona de crecimiento (secreción excesiva de hormona del crecimiento)

Este hecho también lo corrobora Estopà (1999: 227) desde un punto de vista terminológico:

> "La presència d'un article davant del complement és sovint un indici que la unitat no està del tot lexicalitzada."

En la tabla 5 se muestran los patrones que contienen sustantivos como posmodificadores. Puede verse que hay 17 patrones de 20 que tienen uno o más sustantivos como modificadores del núcleo y equivalen al 80,57% (876 ocurrencias). Esto demuestra que el sustantivo es la categoría léxica por excelencia en la posmodificación al menos en este tipo de discurso. Desde un punto de vista sintáctico, se explica que haya más sustantivos que adjetivos como modificadores del núcleo ya que los patrones tienen complementos de nombre, es decir, un sintagma preposicional (SP= Prep+SN). Como se dijo antes, este tipo de discurso, el de las ciencias de la salud, tiende a representar objetos, procesos, fenómenos que son vehiculados por el sustantivo.

Longitud	Patrón	Ejemplo	Frec.	Porc.
3	N Prep N Adj	virus de la inmunodeficiencia humana	343	31,66
3	N Adj Prep N	artrosis degenerativa de la columna	175	16,13
3	N Prep N Prep N	electroforesis en gel de azarosa	118	10,68
4	N Adj Prep N Adj	membrana apical de las células epiteliales	40	3,68
4	N Prep N Adj Prep N	constricción de las arterias coronarias del corazón	36	3,31
3	Adj N Prep N	alto grado de polimorfismo	30	2,76
4	N Adj Prep N Prep N	secreción excesiva de hormona de crecimiento	29	2,67
3	N Prep Adj N	sulfonilurea de alta afinidad	20	1,84
3	N PP Prep N	Oligonucleótidos repetidos en tándem	14	1,29
4	N Prep N Prep N Prep N	Hipocrecimiento por anomalías en genes de los gonosomas	12	1,11
4	N Prep N Prep N Adj	electroforesis en geles de campos pulsantes	11	1,01
3	N Prep N PP	hibridación con sonda marcada	10	0,93
4	N Prep N Adj Adj	inoculación con adenopatías satélites axilares	9	0,83
4	N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74
3	N N Adj	hormona somatomamotropina coriónica	8	0,74
5	N Adj Prep N Prep N Adj	síndrome dismetabólico de sobrecarga de hierro heterocigoto	7	0,64
3	N Adj N	hepatitis vírica C	6	0,55

Tabla 5: Patrones con uno o más sustantivos en la posmodificación.

En la tabla 6 puede observarse que hay 16 patrones de 20 que tienen uno o más adjetivos en la posmodificación y son el 78,82% (855 ocurrencias). Uno de los patrones tiene un adjetivo en posición antepuesta o premodificación. En castellano la presencia de sustantivos y adjetivos es mucho más equilibrada que en inglés, en el cual hay predominio de patrones con sustantivos. Dichos

adjetivos son relaciones a excepción de los adjetivos de gradación o cantidad, e.g. alto, excesivo, etc.

Longitud	Patrón	Ejemplo	Frec.	Porc.
3	N Prep N Adj	virus de la inmunodeficiencia humana	343	31,66
3	N Adj Prep N	artrosis degenerativa de la columna	175	16,13
3	N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
3	N Adj PP	células alveolares descamadas	53	4,88
4	N Adj Prep N Adj	membrana apical de las células epiteliales	40	3,68
4	N Prep N Adj Prep N	constricción de las arterias coronarias del corazón	36	3,31
3	Adj N Prep N	alto grado de polimorfismo	30	2,76
4	N Adj Prep N Prep N	secreción excesiva de hormona de crecimiento	29	2,67
3	N Prep Adj N	sulfonilurea de alta afinidad	20	1,84
4	N Prep N Prep N Adj	electroforesis en geles de campos pulsantes	11	1,01
4	N Prep N Adj Adj	inoculación con adenopatías satélites axilares	9	0,83
4	N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74
3	N N Adj	hormona somatomamotropina coriónica	8	0,74
5	N Adj Prep N Prep N Adj	síndrome dismetabólico de sobrecarga de hierro heterocigoto	7	0,64
3	N Adv Adj	loci altamente polimórficos	7	0,65
3	N Adj N	hepatitis vírica C	6	0,55

Tabla 6: Patrones con uno o más adjetivos como modificadores del núcleo.

En la tabla 7 se presenta el único patrón que contiene adverbios terminados en -*mente* como modificador de un adjetivo y están entre los 20 más frecuentes (N Adv Adj).

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N Adv Adj	loci altamente polimórficos	7	0,65

Tabla 7: Patrones con adverbios.

Como lo propone Kaul (2002: 44), este tipo de adverbios no se reduce simplemente a la interpretación de "modo o manera" sino que vehiculan una cantidad sistemática de significados[53]. De acuerdo con su clasificación muchos de los adverbios terminados en -*mente* del corpus lexicográfico son adverbios de

[53] También plantea la clásica discusión sobre el proceso de formación de este tipo de palabras que va desde la derivación hasta la composición.

punto de vista como se presentan en el ejemplo 12 o adverbios de cantidad o gradación, como en los ejemplos de 13.

12. químicamente indefinido, asintóticamente normal, culturalmente relativista, genéticamente significativa
13. completamente aleatorio, absolutamente sin bias

Es el único ejemplo del corpus de análisis que se alinea en el grupo de los adverbios de cantidad o gradación. En el ejemplo 14, se puede observar que el patrón N Adv Adj es el resultado de una oración pasiva modificada por un adverbio en -*mente* que a su vez modifica a un adjetivo.

14. loci altamente polimórficos (loci que son altamente polimórficos)

En la tabla 8, aparecen los 4 patrones de los 20 más frecuentes que tienen sólo un participio en la premodificación y equivalen al 7,84% (85 ocurrencias). Sintácticamente, los participios pasivos, suelen interpretarse como (que {ha sido/está/puede ser/debe ser} participio de pasado/V) como lo propone Rainer (1999: 4599). Además, Rainer plantea que muchos participios pasivos se emplean en función adjetival y, como tales, parecen pertenecer al dominio de la formación de palabras.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N Adj PP	células alveolares descamadas	53	4,88
3	N PP Prep N	mensajeros controlados por genes	14	1,29
3	N Prep N PP	hibridación con sonda marcada	10	0,93
4	N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74

Tabla 8: Patrones con participios como posmodificadores.

Para observar el predominio de una u otra categoría léxica, se han separado los patrones que contienen sólo adjetivos o sustantivos en los 20 patrones más frecuentes. La tabla 9 contiene los 3 patrones que no tienen sustantivo en la posmodificación y equivalen al 12,26% (133 ocurrencias) del total de la muestra.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
3	N Adj PP	células alveolares descamadas	53	4,88
3	N Adv Adj	loci altamente polimórficos	7	0,65

Tabla 9: Patrones sin sustantivos en la posmodificación.

En la tabla 10 se enseñan los 4 patrones que no tienen adjetivos como modificadores del núcleo y equivalen al 14,01% (154 ocurrencias), pero con participio de pasado (N Prep N PP, N PP Prep N). A diferencia del inglés, en español se mantiene el equilibrio entre los sustantivos y adjetivos como modificadores del núcleo mientras que en inglés el predominio del sustantivo como parte de la premodificación es contundente.

Longitud	Patrón	Ejemplo	Frec.	Porc.
3	N Prep N Prep N	electroforesis en gel de agarosa	118	10,68
3	N PP Prep N	mensajeros controlados por genes	14	1,29
4	N Prep N Prep N Prep N	hipocrecimiento por anomalías en genes de los gonosomas	12	1,11
3	N Prep N PP	hibridación con sonda marcada	10	0,93

Tabla 10: Patrones sin adjetivos en la posmodificación.

En cuanto a las preposiciones, en la tabla 11 los 5 patrones sin preposiciones en la posmodificación representan el 13,55% (147 ocurrencias). Estos patrones normalmente pertenecen a la estructura SAdj o SAdv. Estos datos reflejan que las estructuras que tienden a predominar dentro de la posmodificación, al menos en este tipo de discurso, son sintagmas preposicionales (SPrep) y esto se ve reflejado en la cantidad de preposiciones y sustantivos que acompañan esta estructura.

Longitud	Patrón	Ejemplo	Frecuencia	Porcentaje
3	N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
3	N Adj PP	células alveolares descamadas	53	4,88
3	N N Adj	hormona somatomamotropina coriónica	8	0,74
3	N Adv Adj	loci altamente polimórficos	7	0,65
3	N Adj N	hepatitis vírica C	6	0,55

Tabla 11: patrones sin preposiciones en la posmodificación

5.3.4 Frecuencia de los patrones por longitud

Dependiendo de la longitud, hay patrones que son más regulares que otros al igual que en la lengua general. A continuación, se presentan de mayor a menor extensión los patrones más frecuentes distribuidos por cantidad de tokens.

En la tabla 12, se enseñan los patrones de 5 y 6 tokens. Obsérvese que sólo hay un patrón de 6 tokens con una ocurrencia, lo que no permite hacer ninguna generalización. Posteriormente siguen los patrones de 5 tokens con frecuencias bajas en todo el corpus, pero más frecuentes dentro de los patrones de 5 tokens: N Adj Prep N Prep N Adj (7 ocurrencias), N Adj Prep N Prep N Prep N (4 ocurrencias) y N Prep N Adj Prep N Adj (3 ocurrencias). El resto de 11 patrones tiene 1 ó 2 ocurrencias. En este grupo se observa que el patrón N Adj Prep N Prep N Adj predomina en este grupo, cuya posmodificación está dominada por sintagmas preposicionales. En los ejemplos observados podría decirse que algunas de ellas son unidades libres como lo propone Cabré (1993). Como se ilustra en la tabla 12, algunos ejemplos son unidades libres formadas por unidades de conocimiento y generalmente enlazadas por preposiciones como con, a, para y por (*genoma humano* **con** idéntico *mapa de restricción*), y en algunos casos con determinantes dentro del sintagma (motoneuronas de**l** asta anterior de **la** médula espinal). Cabré (1993) afirma que:

> "En efecto, ante un sintagma terminológico que corresponde a la descripción del contenido de un término, es difícil decidir sin pruebas adicionales si se trata realmente de un término o de una combinación de términos, ya que, aparentemente, entre una combinación libre y una estructura fija no se observa ningún tipo de diferencia."

Esto corrobora lo planteado por Estopà (2006: 258) en cuanto al grado de lexicalización de una unidad terminológica:

> "La presència d'un article davant del complement és sovint un indici que la unitat no està del tot lexicalitzada i, per tant, les estructures en què el complement està introduït per un article determinat percentualment tendeixen a generar més soroll que les estructures en què el complement és indeterminat. En canvi, les estructures $[N[A]_{SAdj}]_{SN}$ i $[[N[A]_{SAdj}]_{SN}[A]_{SAdj}]_{SN}$ generen menys soroll, cosa que no vol dir que en generin poc."

Desde un punto de vista sintáctico, son sintagmas de la estructura SN= SN (SAdj) SPrep SPrep SPrep, en el cual el sintagma preposicional (SPrep) puede reescribirse como SPrep=Prep SN y el SN como N o N Adj. Podría decirse que este tipo de estructuras funciona como expansiones de estructuras de menos longitud, es decir, estructuras con 1 ó 2 SPrep.

Patrón	Ejemplo	Frec.	Porc.
N Prep N Adj Prep N Prep N Adj	actividad de la enzima responsable de la síntesis de óxido nítrico	1	0,09
N Adj Prep N Prep N Adj	síndrome dismetabólico de sobrecarga de hierro heterocigoto	7	0,64
N Adj Prep N Prep N Prep N	visualización directa tras tinción con bromuro de etidio	4	0,37
N Prep N Adj Prep N Adj	motoneuronas del asta anterior de la médula espinal	3	0,28
N Adj PP Prep N Adj	anticuerpos monoclonales ligados a partículas magnéticas	2	0,18
N Prep N Prep N Prep N Prep N	método de delección del cúmulo de hierro en el cuerpo	2	0,18
N Adj Prep Adj N Adj	valores predictivos de los diversos métodos diagnósticos	1	0,09
N Adj Prep Adj N Prep N	genoma humano con idéntico mapa de restricción	1	0,09
N Adj Prep N Adv Adj	fuentes idóneas de linfocitos inmunológicamente activos	1	0,09
N N Prep N Prep N Adj	trisomía X con genes de crecimiento activos	1	0,09
N PP Prep N Prep N Adj	ratas modificadas por medio de ingeniería genética	1	0,09
N Prep Adj N Prep N Adj	endarteritis de pequeños vasos con proliferación endotelial	1	0,09
N Prep N Prep Adj N Prep N	lugares de reconocimiento para distintos factores de transcripción	1	0,09
N Prep N Prep N Adv Adj	confirmación de azoospermia en varones sexualmente maduros	1	0,09
N Prep N Prep N Prep N Adj	vía de transmisión de la señal de modo constante	1	0,09

Tabla 12: Los patrones más frecuentes de 6 y 5 tokens en el corpus de análisis en español.

Los datos de esta extensión muestran que existe una gran variabilidad estructural de este tipo de sintagmas dada la gran cantidad de estructuras (salvo el patrón N Adj Prep N Prep N Adj). Existe una relación de 1,92 sintagmas por

cada patrón. Es decir, que puede haber casi tantos patrones como sintagmas puedan aparecer.

En la tabla 13, se muestran los patrones de 4 tokens (con el núcleo). En total, la muestra contiene 28 patrones de 4 tokens que representan un 16,92% y 181 ocurrencias. Los patrones de 4 tokens agrupan casi la mitad de los patrones de la muestra y, hasta cierto punto, presentan una variabilidad sintáctica importante al tener una relación de 1 patrón por cada 8 sintagmas. El patrón más frecuente es N Adj Prep N Adj con 3,68% y 40 ocurrencias, seguido muy de cerca por los patrones N Prep N Adj Prep N con un 3,31% y 36 ocurrencias y N Adj Prep N Prep N con un 2,67 % y 29 ocurrencias.

Patrón	Ejemplo	Frec.	Porc.
N Adj Prep N Adj	membrana apical de las células epiteliales	40	3,68
N Prep N Adj Prep N	constricción de las arterias coronarias de l corazón	36	3,31
N Adj Prep N Prep N	secreción excesiva de hormona de crecimiento	29	2,67
N Prep N Prep N Prep N	hipocrecimiento por anomalías en genes de los gonosomas	12	1,11
N Prep N Prep N Adj	electroforesis en geles de campos pulsantes	11	1,01
N Prep N Adj Adj	inoculación con adenopatías satélites axilares	9	0,83
N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74
N Adj Prep Adj N	cromatografía líquida de alta resolución	6	0,55
N Adj Adj Adj	poliquistosis renal autosómica recesiva	3	0,28
N Adj Prep N N	terapias regenerativas con células madre	3	0,28
N PP Prep N Adj	metilasas codificadas por los genes kgmA	3	0,56
N Prep Adj N Adj	Hibridación con oligonucleótidos alelo específicos	3	0,28
N PP Prep N Prep N	lactamasas codificadas en plásmidos de enterobacterias	2	0,18
N Prep Adj N Prep N	resistencia a diferentes clases de antibióticos	2	0,18
Adj N Prep N Prep N	escasa especificación de la localización de algunas poblaciones	1	0,09
N Adj Adj N	Hepatitis vírica crónica B	1	0,09
N Adj Adj Prep N	anormalidades genéticas responsables de la tumorigénesis	1	0,09
N Adj Adv Adj	cáncer vesical cistoscópicamente visible	1	0,09
N Adj Adv PP Prep N	enfermedad neuromuscular no ligada al sexo	1	0,09
N Adj PP Adv	bacterias gramnegativas relacionadas serológicamente	1	0,09
N Adj Prep N PP	clonaje posicional de genes mutados	1	0,09
N N Prep N Adj	actividad transferasa en vellosidades curiales	1	0,09
N N Prep N Prep N	actividad proteincinasa sobre residuos de tirosina	1	0,09
N PP Adj Prep N	alelos clonados diferentes del locus	1	0,09
N Prep Adj Adj Prep N	cultivos con medios pobres en folato	1	0,09
N Prep N Adv Adj	azoospermia en varones sexualmente maduros	1	0,09
N Prep N PP Adv	familia de secuencias relacionadas evolutivamente	1	0,09
N Prep N Prep N PP	hemoperfusión con cartucho de carbón activado	1	0,09

Tabla 13: Los patrones más frecuentes de 4 tokens en el corpus de análisis en español.

Puede observarse que, a diferencia de los patrones de 5 tokens, existen patrones que tienen una frecuencia mucho mayor que otros como sucede con los 8 primeros de la tabla, los cuales representan 5 veces más ocurrencias que los otros 20 patrones (151 contra 30) y tienen una alta frecuencia en el corpus. Hay 14 patrones con una ocurrencia (1,26%) y 6 patrones con 2 y 3 ocurrencias (1,76%). Aún así, los patrones de 4 tokens presentan una alta variabilidad sintáctica al tener una media de 6,46 sintagmas por patrón. Salvo los patrones N Adj Adj Adj, N Adj Adj N, N Adj Adv Adj y N Adj PP Adv que pertenecen básicamente a la estructura SN SAdj, el resto de patrones responde a la estructura SN (SPrep) SPrep.

En la tabla 14, se presentan los 17 patrones de 3 tokens. Éstos representan un 80,21% de toda la muestra con 872 ocurrencias. Existe una variabilidad sintáctica de 1 patrón por cada 51,29 sintagmas. Salvo por los patrones N PP Adj, N Prep N N y N N N, todos los patrones de 3 tokens tienen una frecuencia alta, lo que puede revertir en estructuras estables y con tendencia a que estas unidades sean términos.

Patrón	Ejemplo	Frecuencia	Porcentaje
N Prep N Adj	virus de la inmunodeficiencia humana	343	31,66
N Adj Prep N	artrosis degenerativa de la columna	175	16,13
N Prep N Prep N	electroforesis en gel de agarosa	118	10,68
N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
N Adj PP	células alveolares descamadas	53	4,88
Adj N Prep N	alto grado de polimorfismo	30	2,76
N Prep Adj N	sulfonilurea de alta afinidad	20	1,84
N PP Prep N	oligonucleótidos repetidos en tándem	14	1,29
N Prep N PP	hibridación con sonda marcada	10	0,93
N N Adj	hormona somatomamotropina coriónica	8	0,74
N Adv Adj	loci altamente polimórficos	7	0,65
N Adj N	hepatitis vírica C	6	0,55
Adj N Adj	alto peso molecular	5	0,46
N N Prep N	amfotericina B en liposomas	5	0,46
N N N	citocromo c oxidasa	2	0,18
N Prep N N	diabetes de tipo 1	2	0,18
N PP Adj	agua destilada estéril	1	0,09

Tabla 14: Los patrones más frecuentes de 3 tokens en el corpus de análisis en español.

El patrón más frecuente es N Prep N Adj con un 31,66% del total de la muestra y 343 sintagmas, seguido por los patrones N Adj Prep N con un 16,13% y 175 sintagmas y N Adj Prep N con un 10,68% y 118 sintagmas. A su vez son las estructuras más frecuentes en todo el corpus de análisis en español y agrupan el 58,47% de todas las ocurrencias del corpus. Estos tres patrones responden a la estructura SN (SPrep) SPrep.

Al igual que en inglés, se observa que la longitud de los patrones incide directamente en la variabilidad sintáctica, es decir, a mayor extensión, mayor variabilidad, y viceversa.

De igual modo, la productividad de los patrones tiende a disminuir con la extensión al igual que en inglés. Entre más extenso sea un patrón, menos productivo será, entre menos extenso sea, más productivo será. Se observa una tendencia similar que al inglés en cuanto a que los patrones de productividad media de 3 tokens (4,88%) son más productivos que el patrón más productivo de 4 tokens (3,68%) y de 5 tokens (0,64%).

5.3.5 Relaciones de dependencia del corpus de análisis en español

Como se explicó en §3.7, para el análisis de dependencias sintácticas de los patrones en inglés, se seleccionó manualmente una muestra de 8 patrones de los más frecuentes a partir de la muestra del análisis morfosintáctico. Estos 8 patrones representan el 78,30% de todas las ocurrencias del corpus de análisis. Para ello, se seleccionó un 22% de los sintagmas y se distribuyó proporcionalmente de acuerdo con su frecuencia, como se hizo con la muestra sintáctica. Es decir, al patrón más frecuente, le correspondían más sintagmas para el análisis semántico y al patrón menos frecuente se le asignaban menos sintagmas. Por ejemplo, el patrón N Prep N Adj es el más frecuente del corpus y le corresponden 31 sintagmas y los patrones menos frecuentes son N Adj PP y N

Prep Adj N y les corresponden 6 y 5 sintagmas, respectivamente. A su vez, esta muestra se empleó para el análisis semántico del capítulo 7.

En la tabla 15 se lista la frecuencia de dependencias en el conjunto de patrones en español. La relación de dependencia [A [B C]] es la más frecuente en todo el corpus en español como más del 50% de todas las ocurrencias (101) del corpus de análisis.

En esta dependencia el núcleo es modificado al menos por un sintagma preposicional en todos los casos. Este sintagma preposicional rige bien sea al SN o SA, como se ejemplifica en 15.

Dependencia	Frecuencia	Porcentaje
[A [B C]]	101	50,5
[[A B] C]	91	45,5
[[A B] [C D]]	6	3
Ambigua	2	1

Tabla 15: Frecuencias de las dependencias de los patrones en español

15. electroforesis en gel de agarosa, mutaciones de cambio de sentido, tinción con bromuro de etidio, aceptores de puentes de hidrógeno, agenesia de cuerpo calloso, carcinoma de cuello uterino, niveles de ferritina sérica, cDNA de cadena simple, anomalías en el metabolismo del hierro, cambio en la secuencia del DNA, control de la proliferación celular, distribuciones de las frecuencias alélicas, gen de la fibrosis quística, estudios de asociación genética, compromiso de nervios craneanos

A continuación, le sigue la relación de dependencia [[A B] C] con un 45,5% de todas las ocurrencias (91). En esta dependencia, el primer modificador, que en la mayoría de casos es un adjetivo, modifica al núcleo directamente en posición posmodificadora, pero, en algunos casos, en posición premodificadora. Finalmente, este sintagma es modificado por el segundo modificador, como se muestra en los ejemplos de 16.

16. diferentes enzimas de restricción, fuerte desequilibrio de ligamiento, alto grado de homología, candidiasis cutánea generalizada, muerte celular programada, coagulación intravascular diseminada, tumores vesicales superficiales, diabetes mellitus insulinodependiente, fiebre botonosa mediterránea, brazo corto del cromosoma, brazo largo del cromosoma, anticuerpos fijadores del complemento, cáncer de mama en mujeres, secuencia de aminoácidos de SHV-1, cáncer de mama familiar

Por último, aparece la dependencia [[A B] [C D]] con un 3% de todas las ocurrencias (6) para patrones de 4 tokens, como se observa en los ejemplos de 17.

17. asta anterior de la médula espinal, membrana apical de las células epiteliales, alteraciones morfológicas en la biopsia muscular, células epiteliales del túbulo renal, estructura general de las proteínas reguladoras, gel proveniente de los tubos seminíferos

En la tabla 9, se listan las relaciones de dependencia de cada uno de los patrones.

Patrón	Dependencia	Frecuencia	Porcentaje
N Prep N Prep N	[A [[B C]]	19	9,45
N Prep N Prep N	[[A B] C]	7	3,48
N Prep N Prep N	Ambiguo	1	0,5
N Prep N Adj	[A [[B C]]	77	38,31
N Prep N Adj	[[A B] C]	8	3,98
N Prep N Adj	Ambiguo	1	0,5
N Prep Adj N	[A [[B C]]	5	2,49
N Adj Prep N Adj	[[A B] [C D]]	6	2,99
N Adj Prep N	[[A B] C]	44	21,89
N Adj PP	[[A B] C]	6	2,99
N Adj Adj	[[A B] C]	18	8,96
Adj N Prep N	[[A B] C]	8	3,98

Tabla 9. Tipo de dependencia de los patrones de la muestra de análisis.

El patrón N Prep N Prep N tiene dos formas de dependencia: [A [[B C]] con 19 ocurrencias (70,37%) y [[A B] C] con 7 ocurrencias (29,62%). En la primera dependencia, [[A B] C], el sintagma *gen de la hormona del crecimiento*,

el segundo constituyente de la modificación *del crecimiento* modifica directamente al primer constituyente *de la hormona* para formar el sintagma *de la hormona del crecimiento* y, luego éste modifica directamente al núcleo *gen* formar el sintagma *gen de la hormona del crecimiento*, como también puede observarse en los casos del 18.

18. gen de la hormona del crecimiento, hipermotilidad de las articulaciones de las manos, alteración del gen del retinoblastoma, anomalías en el metabolismo de l hierro, cambio en la secuencia de l DNA, hibridación de los extremos de los YACS, Homologías de los alineamientos del DNA, identificación del gen de la miofosforilasa, impedimento de la diferenciación del adipocito, inclusión de las secuencias del plásmido, imprinting en la transmisión de los alelos, transcripción del gen de la prolactina, reparación por escisión de nucleótido, tinción de azul de Perls, electroforesis en gel de agarosa, mutaciones de cambio de sentido, tinción con bromuro de etidio, aceptores de puentes de hidrógeno, replicación en reacciones de PCR

En la segunda dependencia, [[A B] C], el sintagma *cáncer de mama en mujeres*, el primer constituyente de la modificación *de mama* modifica directamente al núcleo *cáncer* para formar un conjunto *cáncer de mama*. Posteriormente, el segundo constituyente (un SP) *en mujeres* modifica al conjunto *cáncer de mama* para formar el sintagma *cáncer de mama en mujeres*, como se aprecia en los casos del 19.

19. cáncer de mama en mujeres, cáncer de mama en varones, secuencia de aminoácidos de CFTR, secuencia de aminoácidos de SHV-1, rutas de señalización de Vav, cristales de cistina en orina, dominio de interacción con Rad51

El patrón N Prep N Adj tiene dos formas de dependencia: [A [[B C]] con 77 ocurrencias (89,53%) y [[A B] C] con 8 ocurrencias (9,30%). En la primera dependencia, [A [[B C]], en el sintagma *células de músculo liso*, el adjetivo liso modifica al sustantivo músculo para formar el sintagma *músculo liso* y este en

su conjunto modifica al núcleo *células* y formar el sintagma *células de músculo liso*, al igual que los ejemplos de 20.

20. agenesia de cuerpo calloso, alteraciones en los parámetros bioquímicos, anemia de células falciformes, análisis de ligamiento genético, análisis de regresión logística, betalactamasas de las bacterias gramnegativas, betalactamasas de las bacterias grampositivas, cDNA de cadena simple, cambio de la flora intestinal, carcinoma de cuello uterino, carcinoma de células transicionales, compromiso de nervios craneanos, concentración de ferritina sérica, concentración de hierro hepático

Obsérvese que en todos los ejemplos el adjetivo que acompaña al segundo sustantivo es relacional formando una especie de compuesto sintagmático que modifica al núcleo. Si fuese un adjetivo calificativo quizá la modificación pudiese afectar directamente al núcleo. La mayoría de estos sustantivos que acompañan al adjetivo relacional son objetos como lo ha observado Demonte (1999: 158-159).

En la segunda dependencia de este patrón, [[A B] C], el sintagma preposicional del primer constituyente de la modificación *hereditario* es el que modifica directamente al núcleo *cáncer* y este conjunto es modificado por el sintagma adjetival *hereditario* para formar el sintagma *cáncer de mama hereditario*, como en los ejemplos de 21.

21. cáncer de mama familiar, cáncer de mama hereditario, factor de crecimiento epidérmico, factores de transcripción específicos, nucleótidos de guanosina radiactivos, vías de señalización intracelulares

El patrón N Prep Adj N tiene la dependencia [A [[B C]] con 5 ocurrencias. Estructuralmente, este patrón puede considerarse una variante del patrón anterior en su primera dependencia, N Prep N Adj. La diferencia radica en que el tipo de adjetivo es calificativo, como se observa en los ejemplos de 22, mientras que los adjetivos del patrón N Prep N Adj son en su mayoría relaciones, como se aprecia en los ejemplos de 20 y 21.

22. familias de alto riesgo, lipoproteínas de baja densidad, rotura de doble cadena, sulfonilurea de alta afinidad, tumores de alto grado

De hecho, si se observan los adjetivos antepuestos o premodificadores del núcleo o de otro sustantivo de todos los patrones, se observa que todos los adjetivos en dicha posición son calificativos, como se ejemplifica en 23.

23. alto, amplio, bajo, buen, cierto, diferente, distinto, diverso, doble, escaso, fuerte, gran, insuficiente, largo, mal, mayor, nuevo, numeroso, sutil, varios

Este tipo de adjetivos dentro de los sintagmas desempeña un papel importante dentro del discurso especializado de acuerdo con Lorente *et al* (2002: 1):

> "La lingüística general ha considerado erróneamente que los adjetivos propios del discurso de especialidad son adjetivos relacionales, mayoritariamente derivados o relacionados formalmente con substantivos, como celular, anatómico, comunicativo. Por el contrario, en este trabajo postulamos que cualquier clase de adjetivo, sea relacional o calificativo, es susceptible de representar y comunicar conocimiento especializado. Parece evidente que tanto el adjetivo fenotípico en el sintagma nominal marcador fenotípico como el adjetivo grande en célula grande, mensajero en RNA mensajero o corto en el sintagma brazo corto del cromosoma contienen y aportan conocimiento especializado."

El patrón N Adj Prep N Adj tiene una sola forma de dependencia [[A B] [C D]] con 6 ocurrencias. En esta dependencia cada adjetivo modifica al sustantivo que acompaña, asta *anterior* y médula *espinal*, y el sintagma preposicional, *de la médula espinal*, modifica luego al sintagma nominal en el cual se haya el núcleo, *asta anterior*. Los ejemplos de 24 siguen este mismo patrón.

24. asta anterior de la médula espinal, membrana apical de las células epiteliales, alteraciones morfológicas en la biopsia muscular, células epiteliales del túbulo renal, estructura general de las proteínas reguladoras

El patrón N Adj Prep N también tiene una sola forma de dependencia [[A B] C] con 44 ocurrencias. En esta dependencia, el primer constituyente de la modificación, el adjetivo *corto*, modifica al núcleo *brazo* y luego el sintagma cromosoma modifica al sintagma nominal *brazo corto* para formar el sintagma *brazo corto del cromosoma*. Los ejemplos de 25 siguen este mismo comportamiento.

25. brazo corto del cromosoma, brazo largo del cromosoma, anticuerpos fijadores del complemento, región codificante del gen, áreas periventriculares del cerebro, expresión genética del colágeno, fase aguda de la enfermedad, arterias coronarias del corazón, artrosis degenerativa de la columna, manifestaciones clínicas de la enfermedad, respuesta clínica al tratamiento, análisis molecular de los genes, bacterias resistentes a los antibióticos, base nitrogenada de las purinas

Los patrones N Adj PP, N Adj Adj y Adj N Prep N tiene una sola forma de dependencia [[A B] C] con 6, 18 y 8 ocurrencias, respectivamente.

En el patrón N Adj PP, el adjetivo *celular* modifica directamente al núcleo *muerte* y el sintagma resultante, *muerte celular* es modificado por el participio de pasado *programada* para formar el sintagma *muerte celular programada*. Los ejemplos de 26 siguen este mismo comportamiento. En los ejemplos puede verse que los adjetivos que acompañan al núcleo son todos relacionales y los sustantivos son objetos o sustantivos deverbales, como lo ha observado Demonte (1999: 159).

26. candidiasis cutánea generalizada, muerte celular programada, coagulación intravascular diseminada, atrofias musculares difusas, crecimiento fetal retardado, campos visuales limitados

En el patrón N Adj Adj, el primer adjetivo *vesicales* modifica al núcleo *tumores* y el segundo adjetivo *superficiales* modifica al nuevo sintagma nominal *tumores vesicales* para formar el sintagma *tumores vesicales superficiales*. Los ejemplos de 27 siguen este mismo comportamiento. Obsérvese que el orden de los adjetivos responde a relacional-relacional (supresor tumoral, adenomatosa familiar, tróficas hipofisarias, etc.) o relacional-calificativo (vesicales superficiales, mieloide crónica, autosómica recesiva, autosómica dominante, renal crónica, etc.), salvo en el caso de *membrana interna mitocondrial* que podría estar también como *membrana mitocondrial interna*.

27. tumores vesicales superficiales, diabetes mellitus insulinodependiente, fiebre botonosa mediterránea, herencia autosómica dominante, leucemia mieloide crónica, sistema nervioso central, genes supresores tumorales, enfermedad autosómica recesiva, forma autosómica recesiva, insuficiencia renal aguda, transmisión autosómica recesiva, gen supresor tumoral, hormonas tróficas hipofisarias, forma autosómica dominante, insuficiencia renal crónica, herencia autosómica recesiva, membrana interna mitocondrial, poliposis adenomatosa familiar

En el patrón Adj N Prep N, el adjetivo premodificador *alto* modifica el núcleo *grado* y este sintagma a su vez, es modificado por *homología* para formar el sintagma *alto grado de homología*. Los ejemplos de 28 siguen este mismo comportamiento. Obsérvese que todos los adjetivos en posición premodificadora son calificativos, como se discutió antes.

28. diferentes enzimas de restricción, fuerte desequilibrio de ligamiento, alto grado de homología, altas tasas de mutación, altas temperaturas de hibridación, alto grado de polimorfismo, alto recuento de leucocitos, diferentes factores de riesgo

5.4 Resultados del corpus lexicográfico de contraste en español

En la tabla 8, se lista cada diccionario, su área temática con su número de entradas y la cantidad de términos de más de tres tokens de longitud y el porcentaje de esta última cantidad para el español. Al igual que en inglés, puede observarse que existe una tendencia similar en cuanto a que entre más extensos sean los diccionarios, menos cantidad de sintagmas de más de tres tokens tienen. Esto quiere decir que entre más grande sea un diccionario, más cantidad de unidades simples puede tener. Por ejemplo, el diccionario Routledge de economía es el diccionario de mayor tamaño con unas 38.000 entradas. Es el diccionario con menor porcentaje de unidades de más de tres tokens (3,92%) que tan sólo representan a 1,491 sintagmas y tiene un elevado número de unidades simples (8.155 ocurrencias) que representan el 21,46 % de todo el diccionario. Como se discutió en §4.4, al revisar algunas de estas unidades simples, puede verse que tienen un carácter terminológico dudoso o sólo funciona como unidades terminológicas generales dentro del diccionario, como se ejemplifica en 29.

29. acción, Alteza, actualizar, pozo, premio, universidad

En cambio, existe la tendencia de que entre menos entradas tenga un diccionario, más unidades de tres tokens de longitud tiene. Por ejemplo, el diccionario IMF tiene 4.500 y es el que tiene el porcentaje más alto de entradas de más de 3 tokens con un 26,31% que representan 1.183 ocurrencias.

Diccionario	Área temática	N.º de entradas	SN de +3 tokens	Porcentaje
Diccionario Mosby	Medicina	31.400	3.848	12,25
Diccionario IFCC	Lab. clínico	4.039	510	12,62
IMF Terminology	Economía	4.500	1.183	26,31
Routledge Dictionary	Finanzas	38.000	1.491	3,92
ISI Multilingual Glossary	Estadística	3.500	883	25,25

Tabla 8: Datos de referencia de cada diccionario en español.

A continuación, se describen los resultados del corpus lexicográfico de contraste en español que se emplea para poder observar las tendencias de extensión, frecuencia de los patrones de más de 3 tokens en diferentes áreas del conocimiento.

5.4.1 Longitud y frecuencia de los SN en los diccionarios en español

En la tabla 9, se presentan los resultados de la longitud de los sintagmas en los diccionarios ordenados de menor a mayor (de tres tokens a más de siete tokens) para el español. Salvo en el caso del diccionario Routledge de economía y finanzas, puede observarse que existe una relación directa entre la extensión del sintagma y la frecuencia de aparición en todos los diccionarios. En el diccionario Routledge tal variación se puede explicar debido a que muchas de las unidades de 3 tokens en este diccionario son nombres propios que no se han tenido en cuenta en este trabajo con lo cual es posible que hubiera mantenido dicha tendencia. Igualmente, es posible que la longitud de 4 tokens sea la que el área de economía y finanzas privilegie desde un punto de vista de la lexicalización. Sin embargo, no se han hecho pruebas para ello ya que no está dentro del alcance de este trabajo. De todos modos, las tendencias presentadas en los otros cuatro diccionarios, muestran dicho predominio en estructuras y sintagmas de 3 tokens.

Diccionario	3 tokens	Porc.	4 tokens	Porc.	5 tokens	Porc.	6 tokens	Porc.	7 tokens+	Porc.
Mosby	3234	84,04	502	13,04	96	2,49	7	0,18	9	0,23
IFCC	359	70,39	112	21,96	23	4,50	9	1,76	7	1,37
IMF	773	65,34	272	22,99	99	8,36	32	2,70	7	0,59
Routledge	91	6,10	1121	75,18	226	15,15	45	3,01	8	0,53
ISI	739	83,69	126	14,26	16	1,81	2	0,22	0	0,0

Tabla 9: Frecuencia por número de tokens del corpus lexicográfico de contraste en español.

En todo el corpus lexicográfico de contraste, los sintagmas de 3 tokens son los más frecuentes (5.196 ocurrencias y un 65,64% en promedio), como se ve en la tabla 9. Por el contrario, los sintagmas de más de 7 tokens son los menos frecuentes (31 ocurrencias y un 0,39% en promedio).

Además, puede verse que los sintagmas de 3 y 4 tokens agrupan el 92,58% de todos los sintagmas del corpus lexicográfico, lo que una vez más confirma los resultados obtenidos por Cartagena (1998) para el español en cuanto a que la extensión de los sintagmas está en el rango de 3 y 4 tokens.

En este corpus lexicográfico sólo el 7,42% representa al resto de sintagmas (de 5 a 8). Desde un punto de vista terminológico, este hecho es muy importante ya que son las unidades que presentan más estabilidad y serían potencialmente más propensas a ser buscadas por el hablante de la lengua. Sin embargo, desde un punto de vista traductivo, este 7,42% de unidades de más de 5 tokens son las que ofrecen más problemas y son las que menos aparecen en los diccionarios para su consulta, por tanto, su bajo nivel de aparición es una desventaja.

5.4.2 Categoría léxica predominante en la modificación de los SN en los diccionarios en español

No existen trabajos en español sobre sintagmación extensa que indiquen qué categoría léxica predomina como modificadora del sustantivo. Sin embargo, debido al potencial que tiene el español para posponer sintagmas preposicionales, puede inferirse que es el sustantivo la categoría que predomina.

Puede observarse en la tabla 10 que la categoría léxica modificadora predominante en todos los diccionarios es el sustantivo. En casi todos los casos, los sustantivos casi duplican a los adjetivos con una media de 38,2% (rango entre 34,49% y 46,67%) mientras que la media de los adjetivos no supera el 21,54% (rango entre 19,33% y 29,3%). A continuación, siguen las preposiciones

con 28,43% (rango entre 22,76% y 32,6%), los determinantes con un 8,25 (rango entre 6,01 y 11,98), los numerales con un 1,34% (rango entre 0,21% y 4,39%), los adverbios con 1,23% (rango entre 0,87% y 2,2%) y los participios de pasado con 0,8% (rango entre 0,03% y 2,67%).

En cuanto al predominio de la categoría gramatical dentro de la premodificación del corpus lexicográfico, hay 34 patrones sin sustantivos en la premodificación de los 445 totales y 122 patrones sin adjetivos, lo que muestra el predominio de los sustantivos como categoría premodificadora. Por otro lado, hay 5 patrones que carecen de sustantivos y adjetivos y 294 patrones con sustantivos y adjetivos a la vez. En cuanto a las otras categorías léxicas abiertas, hay 89 patrones con adverbios y 55 patrones con participio de pasado.

	Mosby		IFCC		IMF		Routledge		ISI	
POS	Frec.	Porc.	Frec.	Porc.	Frec.	Porc.	Frec.	Porc.	Frec.	Porc.
N	4.491	34,74	861	46,67	1.807	34,97	2.808	34,49	1.203	40,17
Adj[54]	3.797	29,3	363	19,67	1.000	19,35	1.629	20,01	579	19,33
PP	4	0,03	0	0	3	0,06	102	1,25	80	2,67
Num[55]	33	0,62	76	4,39	10	0,21	58	0,71	24	0,80
Adv	112	0,87	9	0,49	42	0,81	148	1,82	66	2,20
Prep	3.332	25,78	420	22,76	1.685	32,60	2.647	32,51	862	28,78
V	0	0	0	0	0	0	47	0,58	1	0,03
Det	1.110	8,59	111	6,02	619	11,98	704	8,65	180	6,01

Tabla 10: Categoría léxica predominante en la posmodificación del corpus lexicográfico en español.

[54] Algunos adjetivos se encuentran en posición premodificadora: 25 en el Mosby, 1 en el IFCC, 7 en el IMF, 6 en el Routledge y 4 en el ISI, respectivamente.

[55] Por ajuste de la muestra el porcentaje de esta categoría puede variar ligeramente en los diccionarios Mosby, IMF e IFCC.

5.4.3 Frecuencia de los patrones por aparición en español

En este apartado se analizan los resultados de los patrones del corpus lexicográfico de contraste en español de acuerdo con su frecuencia en todo el corpus y en cada diccionario. De igual modo, se presentan los datos de acuerdo con la extensión del sintagma en cada diccionario.

En su conjunto, el corpus lexicográfico de contraste contiene 445 patrones diferentes y el diccionario con más patrones es el Routledge con 265 y un promedio de 5,62 sintagmas por patrón y el diccionario con menos patrones es el diccionario ISI con 87 patrones y una media de 10,14 sintagmas por patrón. Puede apreciarse en la tabla 11 que los diccionarios con mayor número de entradas (Routledge y Mosby) tienen mayor variabilidad en cuanto a la cantidad de patrones a pesar de que la relación del total de patrones contra total de sintagmas de más de tres tokens pueda ser alta.

Sin embargo, como puede apreciarse en la tabla 12, los siete primeros patrones de cada diccionario representan la mayoría de ocurrencias (rango entre 51,84% y 73,55%) mientras que para el resto corresponde a un número importante de estructuras para unas cuantas ocurrencias. Esto demuestra que también en el corpus lexicográfico de contraste existe una variabilidad sintáctica considerable.

Una vez más, estos datos muestran que la longitud de un sintagma está directamente relacionada con la estabilidad de las estructuras y que hay unas cuantas estructuras (7) que representan a una gran cantidad de sintagmas. Igualmente, una mayor variabilidad sintáctica está relacionada directamente con una (pos)modificación más extensa.

Diccionario	N.º de patrones	SN de +3 tokens	Prom. por patrón
Diccionario Mosby	172	3.848	22,37
Diccionario IFCC	102	510	5,0
IMF Terminology	177	1.183	6,68
Routledge Dictionary	265	1.491	5,62
ISI Multilingual Glossary	87	883	10,14
Total	803 (445 diferentes)	7.401	

Tabla 11: Número de patrones totales del corpus lexicográfico de contraste por diccionario y promedio por patrón.

Como se ilustra en la tabla 12, los patrones más frecuentes en el corpus lexicográfico de contraste son: N Prep N Adj es el más frecuente en los tres diccionarios, además es el segundo más frecuente en uno de ellos y no aparece en el diccionario Routledge; N Prep N Prep N es el segundo más frecuente en dos diccionarios, el tercero más frecuente en dos de ellos y es el número 11 en uno de ellos; N Adj Prep N aparece en cuatro de los diccionarios entre los cinco patrones más frecuentes al igual que N Adj Adj y N Prep N N en tres de ellos. A diferencia del inglés, en español el orden de los patrones varía un poco dentro de los cinco más frecuentes y no conservan el mismo orden de aparición. De todos modos, estas estructuras son las más lexicalizadas y estables de todo el corpus independientemente del área temática y el tamaño del diccionario. El caso excepcional es el diccionario Routledge, el cual no comparte ninguna de estas estructuras con los otros patrones entre los cinco más frecuentes. Sólo el patrón N Prep N Prep N (y la variación N Prep N Prep Num N) que ocupa el puesto 11 es el único de 3 tokens que comparte con el resto del corpus lexicográfico. Se esperaba que este diccionario no siguiera sistemáticamente las tendencias de los otros diccionarios ya que el predominio recae sobre los patrones de 4 tokens, cuyo patrón más frecuente es N Prep N Prep N Adj y en los otros patrones la extensión que predomina es de 3 tokens.

No es fácil saber si esto depende del área temática o de los tipos de datos del diccionario, pero no es posible de establecer en este trabajo, pues se sale del alcance de este estudio.

Es importante destacar otras estructuras del corpus por su frecuencia aparición en los diferentes diccionarios. Entre ellas, pueden destacarse los patrones N N Adj, N N N, N Adj N, N N Prep N y N Adj Prep N Adj presentes dentro de los diccionarios, pero con frecuencias un poco variables.

En cuanto a la distribución de los primeros 20 patrones por número de tokens, puede observarse que, salvo en los diccionarios IMF y en especial el Routledge, los patrones más frecuentes son los de tres tokens, luego siguen los patrones de cuatro tokens y, por último, los de cinco tokens. En cuanto al diccionario Routledge, los patrones de 4 tokens predominan con 15 de 20, seguidos por los de 5 tokens (3 de 20) y luego por los de 3 tokens (2 de 20). En el diccionario IMF, hay un equilibrio entre los patrones de tres y cuatro tokens (9 de 20, en cada caso), seguidos por los patrones de cinco tokens (2 patrones).

Mosby				IFCC				IMF				RD				ISI			
Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%	Tokens	Patrón	Frec.	%
3	N Prep N Adj	743	19,1	3	N Prep N Adj	70	13,6	3	N Prep N Adj	258	21,8	4	N Prep N Prep N Adj	191	12,8	3	N Prep N N	183	20,7
3	N Adj Adj	669	17,2	3	N Prep N Prep N	68	13,2	3	N Prep N Prep N	204	17,2	4	N Adj Prep N Adj	156	10,5	3	N Prep N Adj	131	14,8
3	N Prep N Prep N	437	11,2	3	N N N	63	12,2	3	N Adj Prep N	136	11,5	4	N Adj Prep N Prep N	133	8,92	3	N Prep N Prep N	96	10,9
3	N Adj Prep N	371	9,52	3	N Adj Adj	44	8,54	3	N Adj Adj	65	5,49	4	N Prep N Prep N Prep N	131	8,79	3	N Adj Adj	69	7,81
3	N Prep N N	360	9,25	3	N Adj Prep N	32	6,21	3	N Prep N N	48	4,05	4	N Prep N Adj Prep N	77	5,16	3	N Adj Prep N	66	7,47
3	N N Adj	214	5,49	3	N Prep N N	23	4,46	4	N Prep N Prep N Adj	41	3,46	4	N Prep N Adj Adj	54	3,62	3	N Prep N PP	29	3,28
3	N N N	138	3,55	3	N N Adj	22	4,27	4	N Prep N Prep N Prep N	38	3,21	4	N Prep N Adv Adj	31	2,08	3	N Adj PP	26	2,94
3	N Adj N	83	2,14	4	N N N N	10	1,94	4	N Adj Prep N Prep N	34	2,87	4	N Prep N Prep Adj N	28	1,88	3	N Adv Adj	26	2,94
3	N N Prep N	65	1,67	4	N Adj Prep N Adj	9	1,75	4	N Adj Prep N Adj	30	2,53	4	N Adj Adj Prep N	19	1,27	3	N N Prep N	25	2,83
4	N Prep N Prep N Adj	56	1,44	3	N N Prep N	8	1,55	3	N Prep Adj N	27	2,28	5	N Prep N Prep N Prep N Adj	19	1,27	3	N N Adj	23	2,6
3	N Adv Adj	53	1,36	3	N N Num N	7	1,36	4	N Prep N Adj Prep N	16	1,35	5	N Prep N Adj Prep N Adj	18	1,21	3	N Prep Adj N	21	2,38
4	N Prep N Adj Adj	51	1,32	4	N Prep N N Prep N	7	1,36	3	N Prep N Prep Adj	14	1,18	3	N Prep N Prep N	16	1,07	4	N Prep N Adv Adj	11	1,25
4	N Adj Prep N Adj	49	1,26	4	N Prep N Prep N Prep N	6	1,17	3	N N Prep N	10	0,84	4	N PP Prep N Prep N	15	1,01	3	N Prep N Prep Num N	9	1,02
4	N Adj Adj Adj	45	1,16	3	N Adj N	5	0,97	4	N Prep N Prep N N	10	0,84	3	N Prep N Prep Num N	14	0,94	4	N Prep N N N	9	1,02
3	N Prep Adj N	42	1,08	4	N Prep N Adj Prep N	5	0,97	5	N Prep N Prep N Prep N Prep N	9	0,76	5	N Prep N Adj Prep N Prep N	14	0,94	4	N Prep N Prep N Adj	8	0,91
4	N Adj Prep N Prep N	29	0,74	4	N Prep N Adv Adj	5	0,97	3	N Adj Prep Adj	8	0,67	4	N Adj Prep Adj N	13	0,87	3	N Adj Prep Num N	7	0,79
4	N Prep N Adj Prep N	27	0,69	6	N Adj Prep N Adj N Adj	4	0,77	4	N Adj Prep N N	7	0,59	4	N PP Prep N Adj	13	0,87	4	N Prep N Adj Prep N	7	0,79
4	N Prep N Prep N Prep N	22	0,59	6	N N Adj N N N	4	0,77	5	N Prep N Adj Prep N Prep N	7	0,58	4	N Prep N Prep N N	12	0,8	4	N Prep N Prep N N	7	0,79
4	N Adj Adj Prep N	18	0,46	4	N N Adj Prep N	4	0,78	4	N Prep N Adv Adj	7	0,59	4	N Adj Adj Adj	11	0,74	4	N N N N	6	0,68
4	N Adj Prep N N	14	0,36	3	N Num N Adj	4	0,78	4	N Adv Adj Prep N	6	0,5	4	N Adj Adv Adj	11	0,74	3	N N N	5	0,57

Tabla 12: Los 20 patrones más frecuentes del corpus lexicográfico de contraste en español.

5.4.4 Frecuencia de los patrones por longitud en los diccionarios en inglés

A continuación, se presentan los patrones más comunes distribuidos por longitud de mayor a menor (+7 a 3 tokens). No se mostrarán en las tablas los casos de una y dos ocurrencias, salvo en las extensiones de 6 y 7 tokens en las cuales hay baja frecuencia de estos sintagmas. Sin embargo, se mencionarán en cada caso la cantidad total de patrones de estas frecuencias.

Puede observarse en la tabla 13 que existe una gran variabilidad sintáctica entre los patrones de +7 tokens (23 patrones con una sola ocurrencia). Sólo el diccionario Mosby tiene dos patrones que presentan alguna regularidad (con 3 ocurrencias cada uno). Al igual que en inglés, estos datos muestran que la extensión del patrón está relacionada directamente con una alta variabilidad sintáctica, como se ha planteado en otras partes de esta tesis.

Puede observarse en la tabla 14 que, al igual que los patrones de +7 tokens, existe una gran variabilidad sintáctica entre los patrones de 6 tokens (26 patrones con una sola ocurrencia y 6 con dos ocurrencias). Los patrones más frecuentes son N Adj Prep N Adj N Adj con 6 ocurrencias en dos diccionarios y N N Adj N N N con 4 ocurrencias en un diccionario. Siguen los patrones N Prep N Adj Adj Prep N Adj, N Prep N Adj Prep N N Adj y N Prep N Prep N Prep N Prep N Adj con tres ocurrencias. Es importante resaltar que algunos patrones antes descritos, aparecen sólo en un diccionario con lo que puede decirse que son estructuras idiosincrásicas. De nuevo, los resultados muestran que la extensión está relacionada directamente con una alta variabilidad sintáctica.

	Mosby			IFCC			IMF			Routledge		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
7	N Adj Prep N N Adj N N	3	0,08	N Adj N N N Prep N Adj	1	0,2	N Adj Adv Adj Prep N Prep N Prep N	1	0,08	N Adj Prep N Prep N Prep N Prep N Prep N	1	0,07
7	N Prep N Prep N N Adj N N	3	0,08	N Adj N N N Prep N N	1	0,2	N N Prep N Adj Adj Prep N Adj	1	0,08	N PP Prep N PP Prep N N Adj	1	0,07
7				N N N N N N N	1	0,2	N Prep Adj Adj Prep N Adj	1	0,08	N Prep N Adj Prep N Adj Prep N Adj	1	0,07
7				N N Num N N N N N	1	0,2	N Prep N Adj Prep N Prep N Prep N Adj	1	0,08	N Prep N Adj Prep N Adj Prep N N	1	0,07
7										N Prep N Adv Adj Adj V Prep N	1	0,07
7										N Prep N Prep Adv Adv Adv N V	1	0,07
7										N Prep N Prep N Prep N Prep N Prep N Adj	1	0,07
8				N N N N Adj N N	1	0,2	N Prep N Prep N N Prep N Adj Prep N N	1	0,08			
8				N N N N N N N Adj	1	0,2	N Prep N Prep N Prep N Prep N Adj Prep N Prep N	1	0,08			
8				N N Num N N N Adj N N	1	0,2						
9							N Adj Prep N Prep N N Prep N Prep N Prep	1	0,08	N Prep N Prep N Adj N N V N Prep N	1	0,07

Tabla 13: Los patrones más frecuentes de +7 tokens del corpus lexicográfico de contraste en español.

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
6	N Adj Adj Prep N Adj Adj	2	0,05	N Adj Prep N Adj N Adj	4	0,8	N Prep N Adj Adj Prep N Adj	2	0,17	N Prep N Adj Prep N N Adj	3	0,2	N Prep N N Adv Adj Prep N	1	0,1
6	N Adj Prep N Prep N Prep N Adj	2	0,05	N N Adj N N N	4	0,8	N Prep N Prep N Prep N Prep N Adj	2	0,17	N Adj Prep N Adj N Adj	2	0,13	N Prep N Prep N Prep N Prep N Adj	1	0,1
6	Adj N Prep N Adj Prep N Adj	1	0,03	N N N Adj N N	1	0,2	N Adj Prep N Adj Prep N Adj	1	0,08	N Prep N Prep N Adj Prep N Prep N	2	0,13			
6	N Adj Prep N Adj Prep N N	1	0,03	N Prep N Adj Adj Prep N Adj	1	0,2	N Adj Prep N Prep N Prep Adj N	1	0,08	N Prep N Prep N N Adj Adj	2	0,13			
6	N Prep N Adj Prep N N Prep N	1	0,03	N Prep N Adj Adj Prep N Prep N	1	0,2	N Adj Prep N Prep N Prep N Prep N	1	0,08	N Prep N Prep N Prep N Adj Adj	2	0,13			
6	N Prep N N Adj Prep N Adj	1	0,03				N N N Prep N Prep N N	1	0,08	N Adj Adj PP Prep N Adj	1	0,07			
6	N Prep N Prep N Prep N Adj Prep N	1	0,03				N Prep N Adj Prep N Prep Adj N	1	0,08	N Adj Adj Prep Adv V N	1	0,07			
6							N Prep N Adj Prep N Prep N Prep N	1	0,08	N Adj Adj Prep N Adv Adj	1	0,07			
6							N Prep N N Prep N N Prep N	1	0,08	N Adj Prep N Adj Adj Prep N	1	0,07			
6							N Prep N Prep Adj Prep N Prep N	1	0,08	N Adj Prep N Adj Adv Adj	1	0,07			
6							N Prep N Prep N Adj Prep Adj	1	0,08	N Adj Prep N Adj Prep N N	1	0,07			
6							N Prep N Prep N Adj Prep N Prep N	1	0,08	N Adj Prep N Adj Prep N Prep N	1	0,07			
6							N Prep N Prep N N Prep N Adj	1	0,08	N Adj Prep N PP Prep N Adj	1	0,07			
6							N Prep N Prep N Prep N Adj Prep N	1	0,08	N Adj Prep N Prep N Adj Adj	1	0,07			
6							N Prep N Prep N Prep N N Prep N	1	0,08	N Adj Prep N Prep N Adj Prep N	1	0,07			

Tabla 14: Los patrones más frecuentes de 6 tokens del corpus lexicográfico de contraste en español.

En los patrones de 5 tokens de la tabla 15, puede observarse que aún existe una gran variabilidad sintáctica, pues hay 97 patrones que representan el 39,37% de las ocurrencias (unas 161) de los 121 patrones totales. Sin embargo, hay estructuras claramente más frecuentes que otras y aparecen en varios diccionarios del corpus lexicográfico de contraste. Los 10 primeros corresponden a 172 ocurrencias (42,05%) y los 15 primeros de 209 ocurrencias (51,10%). Hay 23 patrones con +5 de frecuencia y representan de 248 ocurrencias (60,63%) de las 409 ocurrencias en total, es decir, más de la mitad de todas las ocurrencias. Los patrones más frecuentes son N Prep N Prep N Prep N Adj y N Prep N Adj Prep N Adj con un 1,9% (30 ocurrencias) cada uno y aparecen en 3 diccionarios. Finalmente, aparecen los patrones N Prep N Adj Prep N Prep N y N Prep N Prep N Prep N Prep N con 23 ocurrencias (1,57%) y 18 ocurrencias (1,36%), respectivamente.

En cuanto a la exclusividad de patrones, hay 45 patrones que aparecen en más de 2 diccionarios, con lo cual 75 patrones sólo aparecen en un diccionario. De estos 45 patrones, solo 3 aparecen en 4 diccionarios (N Adj Prep N Adj Adj, N Prep N Adj Prep N N y N Prep N Adj Adv Adj) y 13 patrones en 3 diccionarios. Esto muestra que además de la alta variabilidad sintáctica, existe una gran cantidad de estructuras exclusivas de cada diccionario. El diccionario que presenta mayor variabilidad es el Routledge con 85 patrones, de los cuales 47 tienen sólo una ocurrencia. Sin embargo, también es el diccionario con los patrones más frecuentes. Por el contrario, el diccionario con menos patrones es el IFCC con 10 patrones, todos de 1 ó 2 ocurrencias.

Los 10 patrones más frecuentes responden a las estructuras: $(SN\ SA)_{SN}$ $(SP\ (SN\ SA))_{SP}$ $(SP\ (SN\ SA))_{SP}$ y $(SN)_{SN}$ $(SP\ SN)_{SP}$ $(SP\ SN)_{SP}$ $SP\ SN)_{SP}$ $(SP\ (SN\ SA))_{SP}|(SP\ SN)_{SP}$.

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
5	N Adj Prep N Adj Adj	6	0,2	N Adj N N N	2	0,4	N Prep N Prep N Prep N Prep N	9	0,8	N Prep N Prep N Prep N Adj	19	1,3	N Prep N Adj Adv N	2	0,2
5	N Prep N Adj Prep N Adj	6	0,2	N N N N Num N	2	0,4	N Prep N Adj Prep N Prep N	7	0,6	N Prep N Adj Prep N Adj	18	1,2	N Prep N N N N	2	0,2
5	N Adj Adj Prep N N	5	0,1	N N Num N N Prep N	2	0,4	N Prep N Adj Prep N Adj	6	0,5	N Prep N Adj Prep N Prep N	14	0,9	N Adj PP Prep N Adj	1	0,1
5	N Adj Prep N Prep N Adj	5	0,1	N Prep N Adv Adj Prep N	2	0,4	N Prep N Prep N Prep N Adj	6	0,5	N Adj Prep N Prep N Prep N	10	0,7	N Adj Prep N Adj Adj	1	0,1
5	N Prep N Adj Adj Prep N	5	0,1	N Adj Prep N N N	1	0,2	N Prep N Adj Prep N N	4	0,3	N Adj Prep N Prep N Adj	9	0,6	N Adj Prep N Adv Adj	1	0,1
5	N Prep N Prep N Prep N Adj	5	0,1	N N N Prep N Adj	1	0,2	N Adj Prep N Prep N Prep N	3	0,3	N Prep N Prep N Prep N Prep N	9	0,6	N Adj Prep N Prep Adj N	1	0,1
5	N Adj Adj Prep N Adj	4	0,1	N Num N N N N	1	0,2	N Prep N Prep N N Prep N	3	0,3	N Prep N Prep N Adj Prep N	6	0,4	N N Adv N Adj	1	0,1
5	N Adj Prep N Adj Prep N	3	0,1	N Prep N Adj Adj N	1	0,2	N Adj Prep N N Prep N	2	0,2	N Adj Adj Prep N Prep N	5	0,3	N Prep N Adj Adv Adj	1	0,1
5	N Prep N N Prep N Adj	3	0,1	N Prep N Adj N N	1	0,2	N Adj Prep N Prep N Adj	2	0,2	N Prep N Adj Prep Adj N	5	0,3	N Prep N Adj Prep N N	1	0,1
5	N Prep N Prep N Prep Adj N	3	0,1	N Prep N Adj Prep Adj N	1	0,2	N Adv Adj Prep N Prep N	2	0,2	N Prep N Adj Prep N N	5	0,3	N Prep N N N Adj	1	0,1
5	N Adj Adj Adj Prep N	2	0,1				N Prep N Adj Adv Adj	2	0,2	N Prep N Prep N Prep Adj N	5	0,3	N Prep N PP Prep N Prep N	1	0,1
5	N Adj Adj Prep N Prep N	2	0,1				N Prep N N Prep N Adj	2	0,2	N Adj Prep N Adj Adj	4	0,3	N Prep N Prep Adj N Prep Adj	1	0,1
5	N Adj N Adv Adj	2	0,1				N Prep N Prep N Adj Adj	2	0,2	N Adj Prep N Adj Prep N	4	0,3	N Prep N Prep N Adv Adj	1	0,1
5	N Adj Prep N Prep N Prep N	2	0,1				N Prep N Prep N N Adj	2	0,2	N Adj Prep N Adv Prep N	4	0,3			
5	N Prep N Adj Prep N Prep N	2	0,1				N Adj Adj Prep N Adj	1	0,1	N Adj Prep N N Adj	4	0,3			
5	N Prep N Prep N Adj Adj	2	0,1				N Adj Adj Prep N N	1	0,1	N Prep N Prep N Adj N	4	0,3			
5	Adj N Prep N Adj Adj	1	0				N Adj Adv Adj Prep N	1	0,1	N Adj Adj Prep N Adj	3	0,2			
5	Adj N Prep N Prep N Adj	1	0				N Adj Prep N Adj Adj	1	0,1	N Adj Adj Prep N N	3	0,2			
5	Adj N Prep N Prep N	1	0				N Adj Prep N Adj Prep	1	0,1	N Adj N Prep N Prep N	3	0,2			

	Prep Adj						N								
5	Adj N Prep N Prep N Prep N	1	0				N Adj Prep N N N	1	0,1	N Adj Prep N Prep Adj N	3	0,2			
5	N Adj Adj Adj Adj	1	0				N Adj Prep N Prep N N	1	0,1	N Adj Prep N Prep N N	3	0,2			
5	N Adj Adj Adv Adj	1	0				N Adv Adj Prep N Adj	1	0,1	N Adv Adj Prep N Adj	3	0,2			
5	N Adj Adj Adv N	1	0				N Adv Adj Prep N N	1	0,1	N Adv Adj Prep N N	3	0,2			
5	N Adj Prep N N Adj	1	0				N N Prep N Adj Prep N	1	0,1	N Prep Adj N Prep N Adj	3	0,2			
5	N N Adj N Prep N	1	0				N N Prep N N N	1	0,1	N Prep N Prep N Adv Adj	3	0,2			
5	N N Adj Prep N Prep N	1	0				N N Prep N Prep N Prep N	1	0,1	N Prep N Prep N N N	3	0,2			

Tabla 15: Los patrones más frecuentes de 5 tokens del corpus lexicográfico de contraste en español.

En los patrones de 4 tokens de la tabla 16, puede observarse que existe mucha menos variabilidad sintáctica ya que 20 patrones representan el 79,39% de las ocurrencias (unas 1.576) del total de 1.985. A diferencia de los patrones de 5 tokens, los patrones de 4 tokens presentan estructuras claramente más frecuentes y muchas de ellas aparecen en todos los diccionarios del corpus lexicográfico de contraste. Los 10 primeros representan a 1.316 ocurrencias (66,29%) y los 20 primeros de 1.576 ocurrencias (79,39%). Hay 43 patrones con +5 de frecuencia y representan a 1.804 ocurrencias (90,88%) de las 1.985 ocurrencias en total, es decir, la mayoría de ocurrencias. Los patrones más frecuentes son N Prep N Prep N Adj con 299 ocurrencias (15,06%), N Adj Prep N Prep N con 202 ocurrencias (10,17%) y N Prep N Prep N Prep N con 201 ocurrencias (10,12%) y aparecen en los 5 diccionarios. Luego, siguen los patrones N Adj Prep N Adj, N Prep N Adj Prep N y N Prep N Adj Adj con 156 ocurrencias (7,85%), 132 ocurrencias (6,64%) y 120 ocurrencias (6,0%), respectivamente.

En cuanto a la exclusividad de patrones, hay 51 patrones que aparecen en más de 2 diccionarios, con lo cual 100 patrones sólo aparecen en un diccionario. De estos 51 patrones, 20 patrones aparecen en los 5 diccionarios con 1.525 ocurrencias, 6 patrones aparecen en los 4 diccionarios con 85 ocurrencias, 8 patrones aparecen en los 3 diccionarios con 125 ocurrencias y 17 patrones aparecen en los 2 diccionarios con 76 ocurrencias. Esto muestra que existe menos variabilidad sintáctica que en los otros casos, aunque existe una gran cantidad de estructuras exclusivas de cada diccionario. Es importante destacar que esta variabilidad sería muy baja si el diccionario Routledge no tuviera 61 patrones de 1 ó 2 ocurrencias. Así, el diccionario que presenta mayor variabilidad es el Routledge con 104 patrones, de los cuales 44 tienen sólo una ocurrencia. Sin embargo, también es el diccionario con los patrones más frecuentes como sucede con los patrones de 5 tokens. Por el contrario, el diccionario con menos patrones es el IFCC con 41 patrones, de cuales 26 son de 1 ó 2 ocurrencias de los que se infiere que también tiene mucha exclusividad de

patrones. El diccionario con menos exclusividad de patrones es el Mosby con sólo 14 patrones de 1 ocurrencia.

	Mosby			IFCC[56]			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
4	N Prep N Prep N Adj	56	1,44	N N N N	10	1,9	N Prep N Prep N Adj	41	3,46	N Prep N Prep N Adj	191	12,8	N Prep N Adv Adj	11	1,3
4	N Prep N Adj Adj	51	1,32	N Adj Prep N Adj	9	1,8	N Prep N Prep N Prep N	38	3,21	N Adj Prep N Adj	156	10,5	N Prep N N N	9	1
4	N Adj Prep N Adj	49	1,26	N Prep N N Prep N	7	1,4	N Adj Prep N Prep N	34	2,87	N Adj Prep N Prep N	133	8,92	N Prep N Prep N Adj	8	0,9
4	N Adj Adj Adj	45	1,16	N Prep N Prep N Prep N	6	1,2	N Adj Prep N Adj	30	2,53	N Prep N Prep N Prep N	131	8,79	N Prep N Adj Prep N	7	0,8
4	N Adj Prep N Prep N	29	0,74	N Prep N Adj Prep N	5	1	N Prep N Adj Prep N	16	1,35	N Prep N Adj Prep N	77	5,16	N Prep N Prep N N	7	0,8
4	N Prep N Adj Prep N	27	0,69	N Prep N Adv Adj	5	1	N Prep N Prep N N	10	0,84	N Prep N Adj Adj	54	3,62	N N N N	6	0,7
4	N Prep N Prep N Prep N	22	0,59	N N Adj Prep N	4	0,8	N Adj Prep N N	7	0,59	N Prep N Adv Adj	31	2,08	N Adj Prep Adj N	5	0,6
4	N Adj Adj Prep N	18	0,46	N Prep N Adj Adj	4	0,8	N Prep N Adv Adj	7	0,59	N Prep N Prep Adj N	28	1,88	N Prep N Adj Adj	5	0,6
4	N Adj Prep N N	14	0,36	N Prep N Prep Adj N	4	0,8	N Adv Adj Prep N	6	0,5	N Adj Adj Prep N	19	1,27	N Prep N Prep Adj N	5	0,6
4	N Prep N N Adj	13	0,34	N Adj Num N Adj	3	0,6	N Prep N Adj Adj	6	0,51	N PP Prep N Prep N	15	1,01	N Prep N Prep N PP	5	0,6
4	N Prep N N N	13	0,33	N Adj Prep N Prep N	3	0,6	N Prep N Prep Adj N	6	0,51	N Adj Prep Adj N	13	0,87	N Adj Adv Adj	4	0,5
4	N Adj Adv Adj	10	0,26	N Prep N N N	3	0,6	N Adj Adj Prep N	5	0,42	N PP Prep N Adj	13	0,87	N Prep N N Prep N	4	0,5
4	N N N Adj	9	0,24	N Prep N Prep N Adj	3	0,6	N Adj Adv Adj	5	0,42	N Prep N Prep N N	12	0,8	N Prep N Prep N Prep N	4	0,5
4	N Prep N Prep N N	9	0,23	N Prep N Prep N N	3	0,6	N Adj Prep Adj N	5	0,42	N Adj Adj Adj	11	0,74	N Adj Adv PP	3	0,3
4	N Adj N Adj	8	0,21	Num N N N N	3	0,6	N Prep N N Adj	4	0,34	N Adj Adv Adj	11	0,74	N Adj Prep N N	3	0,3
4	N N N N	7	0,18	N Adj N N	2	0,4	N Adj N N	3	0,25	N Prep N PP Prep N	11	0,74	N Adj Prep N Prep N	3	0,3
4	N Prep N Adj N	7	0,19	N N Adj Adj	2	0,4	N Adj Prep Adj N Prep N	3	0,24	N Prep N Adj N	10	0,67	N Prep N N Adj	3	0,3
4	N N Adj Adj	6	0,15	N N N Adj	2	0,4	N N Prep N Adj	3	0,25	N Prep N Adv Prep N	10	0,67	N Adj Adj Adj	2	0,2
4	N N Adj Prep N	6	0,15	N N N N Adj	2	0,4	N N Prep N Prep N	3	0,25	N Adj Prep N N	9	0,6	N Adj Prep N Adj	2	0,2
4	N N Prep N Adj	6	0,15	N Adj Adj N	1	0,2	N Prep N Adj Prep Adj	3	0,25	N N Prep N Adj	8	0,54	N Adv Adv Adj	2	0,2
4	N Prep N N Prep N	6	0,15	N Adj Adj Prep N	1	0,2	N Prep N N N	3	0,25	N Prep N Adj Prep Adj	8	0,54	N N Adj Adj	2	0,2
4	N Adj Prep Adv Adj	5	0,13	N Adj Adv Adj	1	0,2	N Prep Adj Prep N Adj	2	0,16	N Adv Adj Prep N	7	0,47	N N N Adj	2	0,2
4	N Prep N Adv Adj	5	0,13	N Adj N Prep N	1	0,2	N Prep N Prep Adj Prep N	2	0,17	N Adj N Prep N	6	0,4	N N Prep N Adj	2	0,2
4	N Adj Adj N	3	0,08	N Adj Prep Adj N	1	0,2	Adj N Adj Prep N	1	0,08	N N Prep N Prep N	6	0,4	N Prep N Adv N	2	0,2
4	N Adj N Prep N	3	0,08	N Adj Prep N N	1	0,2	Adj N Prep N N	1	0,08	N Prep Adj N Adj	6	0,4	Adj N Adj PP	1	0,1

[56] El diccionario de laboratorio clínico tiene muchos términos que son nomenclaturas y por esta razón muchos de los patrones y datos en general no concuerdan con los de los otros diccionarios.

4	N Adj Prep Adj N	3	0,08	N N Adj N Prep N	1	0,2	N Adj Adv Prep N	1	0,08	N Adj N Adj	5	0,34	N Adj Adj N	1	0,1
4	N N Adv Adj	3	0,08	N N Adj Num N Adj	1	0,2	N Adj N Adj	1	0,08	N Adv Prep N Adj	5	0,34	N Adj Adj PP	1	0,1
4	N N Prep N Prep N	3	0,08	N N N Prep N	1	0,2	N Adj N Prep N	1	0,08	N Adv Prep N Prep N	5	0,34	N Adj Adj Prep N	1	0,1
4	Adj N Prep N Prep N	2	0,05	N N Num N Adj	1	0,2	N Adj Prep N Prep N Num	1	0,08	N Prep N Adv PP	5	0,34	N Adj N Adj	1	0,1
4	N Adj Adv N	2	0,05	N N Num N N	1	0,2	N Adv Adj Adj	1	0,08	N Prep N N N	5	0,34	N Adj N N	1	0,1
4	N Adj N N	2	0,05	N N Prep N Prep N	1	0,2	N Adv Adj Prep Adj	1	0,08	N Adj Prep Adj Prep N	4	0,27	N Adj N Prep N	1	0,1
4	N Adj N N Num	2	0,05	N Num Adj Num N Adj	1	0,2	N N Adj Adj	1	0,08	N Prep N N Prep N	4	0,27	N Adv Adj Prep N	1	0,1
4	N N Adj N	2	0,05	N Num N Adj Adj	1	0,2	N N N Adj	1	0,08	N Prep N Prep N Prep V	4	0,27	N N Adj PP	1	0,1
4	N PP Prep N Adj	2	0,05	N Num N N N	1	0,2	N N N N	1	0,08	Adj N Adj Prep N	3	0,2	N N Adj Prep N	1	0,1
4	N Prep Adj Adj Prep N	2	0,05	N Num Prep N N N	1	0,2	N N Prep N Prep Adj N	1	0,08	N Adj N N	3	0,2	N N Prep Adj N	1	0,1
4	N Prep N Prep Adj N	2	0,05	N Prep N Adj N	1	0,2	N Prep Adj N Prep N	1	0,08	N Adj Prep N Prep V	3	0,2	N N Prep N PP	1	0,1
4	Adj N Adj Prep N	1	0,03	N Prep N N N N	1	0,2	N Prep Adv Adj N	1	0,08	N Adj Prep PP Prep N	3	0,2	N N Prep N Prep N	1	0,1
4	Adj N Prep N Adj	1	0,03	N Prep N Prep Adv Adj N	1	0,2	N Prep Adv N Prep N	1	0,08	N N N Adj	3	0,2	N Prep Adj N Adj	1	0,1
4	N Adj N Prep Num N	1	0,03	Num Adj Num N N Adj	1	0,2	N Prep N Adj Prep Prep N	1	0,08	N PP Prep Adj N	3	0,2	N Prep Adj N N	1	0,1
4	N Adv Prep N N	1	0,03	Num N Adj N Num N	1	0,2	N Prep N Adv N	1	0,08	N Prep Adj N Prep N	3	0,2	N Prep N Adj PP	1	0,1
4	N N N Prep N	1	0,03	Num N N N N	1	0,2	N Prep N N Prep N	1	0,08	N Prep Adj Prep N Adj	3	0,2	N Prep N Adv PP	1	0,1
4	N N Prep Adj Adj	1	0,03				N Prep N Prep N Prep N Num	1	0,08	N Prep N Adj Prep Num N	3	0,2	N Prep N N N Num	1	0,1
4	N N Prep Adv Adj	1	0,03							N Prep N Prep V N	3	0,2	N Prep N N N Prep	1	0,1

Tabla 16: Los patrones más frecuentes de 4 tokens del corpus lexicográfico de contraste en español.

En los patrones de 3 tokens de la tabla 17, puede observarse que son los sintagmas con menos variabilidad y exclusividad de sintagmas por diccionario. En los patrones de 3 tokens se reduce la variabilidad sintáctica ya que 20 patrones representan el 97,20% de las ocurrencias (unas 5.147) del total de 5.325. En el caso de los patrones de 3 tokens, se presentan estructuras con frecuencias altas y muchas de ellas aparecen en todos los diccionarios del corpus lexicográfico de contraste. Los 10 primeros representan a 4.877 ocurrencias (92,10%) y los 20 primeros de 5.147 ocurrencias (97,20%). Hay 27 patrones con +5 de frecuencia y corresponden a 52,02 ocurrencias (98,24%) de las 5.295 ocurrencias en total, es decir, casi la totalidad de las ocurrencias. Los patrones más frecuentes son N Prep N Adj con 1.209 ocurrencias (22,83%), N Adj Adj con 848 ocurrencias (16,01%) y N Prep N Prep N con 821 ocurrencias (15,50%) y aparecen en los 5 diccionarios al igual que los patrones N Prep N N con 617 ocurrencias (11,65%) y N Adj Prep N con 611 ocurrencias (11,53%).

En cuanto a la exclusividad de patrones, hay 29 patrones que aparecen en más de 2 diccionarios, con lo cual 52 patrones sólo aparecen en un diccionario. De estos 29 patrones, 9 patrones aparecen en los 5 diccionarios con 4.463 ocurrencias, 3 patrones aparecen en los 4 diccionarios con 449 ocurrencias, 110 patrones aparecen en los 3 diccionarios con 198 ocurrencias y 7 patrones aparecen en los 2 diccionarios con 36 ocurrencias. Al igual que en los patrones de 4 tokens, existe menos variabilidad sintáctica, aunque existe una gran cantidad de estructuras exclusivas de cada diccionario lo que explica los 52 patrones. Es importante destacar que esta variabilidad sería muy baja si los diccionarios Mosby, IMF y Routledge no tuvieran 17, 15 y 19 patrones de 1 ó 2 ocurrencias, respectivamente. De igual modo, los otros dos diccionarios tienen unos 10 patrones, cada uno de 1 ó 2 ocurrencias. Así, el diccionario que presenta mayor variabilidad es el Routledge con 29 patrones y un promedio de 3,13 sintagmas por cada patrón (91 totales), de los cuales 19 patrones tienen 1 ó 2 ocurrencias. Por el contrario, el diccionario con menos variabilidad es el Mosby con 40 patrones y un promedio de 81,92 sintagmas por cada patrón (3.277 totales), de los cuales 17 patrones tienen 1 ó 2 ocurrencias. Puede observarse que

a pesar de tener más patrones que los otros diccionarios, cada uno de sus patrones tiene una frecuencia alta. El diccionario IFCC tiene un promedio de 13,46 por cada patrón, el diccionario IMF tiene un promedio de 30,03 por cada patrón y el diccionario ISI promedio de 27,37 por cada patrón.

	Mosby			IFCC			IMF			Routledge			ISI		
Tokens	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%	Patrón	Frec.	%
3	N Prep N Adj	743	19,1	N Prep N Adj	70	14	N Prep N Adj	258	21,79	N Prep N Prep N	16	1,07	N Prep N N	183	21
3	N Adj Adj	669	17,2	N Prep N Prep N	68	13	N Prep N Prep N	204	17,23	N Prep N Prep Num N	14	0,94	N Prep N Adj	131	15
3	N Prep N Prep N	437	11,2	N N N	63	12	N Adj Prep N	136	11,49	N Adj Prep Num N	9	0,6	N Prep N Prep N	96	11
3	N Adj Prep N	371	9,52	N Adj Adj	44	8,5	N Adj Adj	65	5,49	N Prep N Adj	8	0,54	N Adj Adj	69	7,8
3	N Prep N N	360	9,25	N Adj Prep N	32	6,2	N Prep N N	48	4,05	N Adj Prep N	6	0,4	N Adj Prep N	66	7,5
3	N N Adj	214	5,49	N Prep N N	23	4,5	N Prep Adj N	27	2,28	Num N Prep N Prep N	4	0,27	N Prep N PP	29	3,3
3	N N N	138	3,55	N N Adj	22	4,3	N Prep N Prep Adj	14	1,18	N Adj Prep Adj	3	0,2	N Adj PP	26	2,9
3	N Adj N	83	2,14	N N Prep N	8	1,6	N N Prep N	10	0,84	N N N Num	3	0,2	N Adv Adj	26	2,9
3	N N Prep N	65	1,67	N N Num N	7	1,4	N Adj Prep Adj	8	0,67	N PP Prep N	3	0,2	N N Prep N	25	2,8
3	N Adv Adj	53	1,36	N Adj N	5	1	N N N	5	0,42	N Prep N N	3	0,2	N N Adj	23	2,6
3	N Prep Adj N	42	1,08	N Num N Adj	4	0,8	N Adj N	4	0,34	N Adj Prep Prep N	2	0,13	N Prep Adj N	21	2,4
3	N Prep Adj Adj	13	0,33	N Prep N N Num	4	0,8	N Prep Adj Adj	4	0,34	N N N	2	0,13	N Prep N Prep Num N	9	1
3	N Prep Adv Adj	10	0,26	N Num Num N Adj	3	0,6	N Prep Adj Prep N	4	0,33	N Prep Prep N Adj	2	0,13	N Adj Prep Num N	7	0,8
3	N Prep Adj Prep N	9	0,23	Num N N N	3	0,6	N Adv PP	3	0,25	Adj N Prep Adv	1	0,07	N N N	5	0,6
3	N Prep N Prep Adj	9	0,23	N Adj Prep Num N	2	0,4	N Prep N N Num	3	0,25	Adj N Prep N	1	0,07	N Adj N	4	0,5
3	Adj N Prep N	8	0,21	N N Prep N Num	2	0,4	Adj N Adj	2	0,17	N Adj Adj	1	0,07	N N PP	4	0,5
3	N Adj Prep Adj	8	0,21	N Num N N	2	0,4	Adj N Prep N	2	0,17	N Adj Prep N Prep Num	1	0,07	N Adv PP	3	0,3
3	Adj N Adj	7	0,18	N Num N Num Adj	2	0,4	N Adj Adv	2	0,17	N Adv Prep N	1	0,07	N Adv Prep N	2	0,2
3	N Adv N	4	0,1	N Num Num Num Num Adj N	2	0,4	N Adj Prep Num N	2	0,17	N Prep N Adj Prep	1	0,07	N Prep N Num Prep Num N	2	0,2
3	N N N Num	4	0,1	N Prep Adj N	2	0,4	N N Adj	2	0,17	N Prep N Adj Prep Num	1	0,07	Adj N Adj	1	0,1
3	N Adj Adj Prep Num	3	0,08	Adj N Prep N	1	0,2	N Num Prep N N	2	0,17	N Prep N Prep Num Prep N	1	0,07	Adj N Prep N	1	0,1
3	N Adv Prep N	3	0,08	N Adj Num N	1	0,2	Adj N N	1	0,08	N Prep N Prep V	1	0,07	Adj N Prep Num N	1	0,1
3	N Prep N Prep Num N	3	0,08	N N N Num	1	0,2	N Adj Prep Adv	1	0,08	N Prep Num N Adj	1	0,07	N N Prep Num N	1	0,1

Tabla 17: Los patrones más frecuentes de 3 tokens del corpus lexicográfico de contraste en español.

5.5 Contraste de resultados entre el corpus de análisis y el corpus lexicográfico en español

A continuación, se comparan los resultados obtenidos en el corpus de análisis en español y el corpus lexicográfico de contraste en cuanto a la distribución por longitud, la categoría léxica predominante en la premodificación, la frecuencia de patrones por aparición y la frecuencia de los patrones de acuerdo con la longitud.

5.5.1 Distribución de acuerdo con la longitud

Como puede verse en la tabla 18, la comparación de los datos de ambos corpus permite establecer que efectivamente las tendencias presentadas en ambos casos muestran que las estructuras más estables están directamente relacionadas con una menor extensión. A diferencia de la igualdad de porcentajes encontrada en los corpus en inglés, en español existe una diferencia más marcada entre la extensión de los patrones del corpus de análisis que entre la extensión de los patrones del corpus lexicográfico. Aún así, la preferencia de los diccionarios por lexicalizar estructuras más cortas sigue la tendencia antes revisada aunque menos pronunciada.

	Corpus		Diccionarios	
N.º tokens	Frecuencia	Porcentaje	Frecuencia	Porcentaje
3	872	80,66	5.196	65,64
4	181	16,74	2.133	26,94
5	27	2,5	460	5,81
6	1	0,09	95	1,2
7	0	0	31	0,39
Total	1.081	100	7.915	100

Tabla 18: Frecuencia por número de tokens entre el corpus de análisis y el corpus lexicográfico en español.

Así, los patrones de 3 tokens en el corpus tenderán a ser las estructuras más estables, y los sintagmas que tienen estas estructuras serían candidatos a términos.

5.5.2 Categoría léxica predominante y aspectos morfológicos

Al igual que en el corpus de análisis, los sustantivos en el corpus lexicográfico predominan como categoría modificadora del núcleo en los sintagmas, como puede verse en la tabla 19. Sin embargo, el porcentaje es mayor ya que casi duplican a los adjetivos mientras que en el corpus de análisis los sustantivos sólo superan a los adjetivos, en sólo un 8,7%. De nuevo se corrobora que la tendencia léxica no sólo del discurso especializado sino de los diccionarios científico-técnicos.

Así, las tendencias léxicas del discurso especializado en el uso del sustantivo como categoría léxica preferida no es dependiente del área temática sino del discurso especializado.

	Corpus	Diccionarios
Cat. léxica	Porcentaje	Porcentaje
N (sin núcleos)	31,78	38,2
Adj	23,05	21,54
PP	4,98	0,8
Adv	2,49	1,23
Prep	28,35	28,48
otras	9,35	9,71

Tabla 19: Comparación de la categoría léxica predominante entre el corpus de análisis y el corpus lexicográfico en español.

Los sustantivos representan casi la mitad de las unidades léxicas que aparecen en ambos corpus. Morfológicamente, la mayoría de sustantivos terminan en *–ción*, y sus alomorfos con 227/129 sustantivos en el corpus de análisis y 1.440/1.126 en el corpus lexicográfico en los núcleos y la modificación,

respectivamente. Gallegos (2003: 43) también reporta este sufijo como el más productivo en el español científico sino también técnico y general aunque su frecuencia en texto científico es mucha más alta que en los otros dos tipos. El orden de frecuencia de los sufijos varía en ambos corpus tanto como núcleo como modificador. El segundo sufijo más común en ambos corpus, aunque solo en los núcleos, es *–dad* y sus variantes con 52/310, respectivamente y no reportado por Gallegos (2003). Mientras que este mismo sufijo es el cuarto y tercero más frecuente en la modificación (29/235, respectivamente). A pesar de que no siguen el mismo orden de frecuencia los diez primeros sufijos presentes en el corpus de análisis están presentes en los diez o doce primeros del otro corpus. Entre ellos, pueden mencionarse los terminados en *-ía, -or, -dor, -ura, -ido -m(i)ento, -ncia,* y *–ado*, estos tres últimos también reportados por Gallegos (2003) entre los 10 más productivos del español científico. Como bien lo plantea Varela (2005: 49) estos sufijos son deverbales al igual que *–ción*; se derivan de verbos, como se aprecia en los siguientes casos de los ejemplos 30, 31 y 32.

30. administración, disregulación, hibridación, avulsión, betaoxidación, fosforilación, inoculación, instilación, luxación, monitorización, osificación, sobreexpresión, subluxación, tinción.

31. adherencia, contingencia, dependencia, discordancia, discrepancia, experiencia, negligencia, penetrancia, prevalencia, resistencia, resonancia, suficiencia, transferencia, tumescencia.

32. hipocrecimiento, abotargamiento, acaparamiento, agotamiento, ajustamiento, amamantamiento, asentimiento, atrapamiento, financiamiento, ligamento, procedimiento, reforzamiento, requerimiento, taponamiento.

Como lo apuntan Lacuesta y Bustos (1999: 4511) todos los sufijos anteriores se enmarcan en significado de “acción” y por tanto reflejan un conjunto de propiedades semánticas asociadas a la nominalización.

Es difícil establecer si esta productividad es propia del discurso científico o es igual en la lengua española en general, salvo en los cuatros sufijos

mencionado por Gallegos (2003) ya que no se conocen estudios comparativos de este tipo en corpus de medio o gran tamaño[57]. Pero si se comparan con lo planteado para el inglés en §4.5.2, los tres primeros presentan el mismo orden de productividad observado por Biber *et al* (1999: 322-323) para el discurso académico.

De los datos se deduce que la nominalización es el recurso más eficiente dentro del discurso especializado y que algunas formas son típicas de los "lenguajes especializados" (Gallegos 2003: 37). Como se ha dicho antes, su uso se justifica pragmáticamente por los objetivos que se persiguen en la ciencia: universalidad, revisabilidad y verificabilidad (Vivanco 2005: 19).

Posteriormente, siguen los adjetivos con un 23,05% y de las unidades léxicas en el corpus de análisis y un 21,54 en el corpus lexicográfico. No existe una diferencia importante entre ambos corpus. En cuanto a los adjetivos, la mayoría de estos se caracterizan morfológicamente por terminar en dos sufijos con base nominal *–ico* (196/1.085 ocurrencias en ambos corpus, respectivamente) y *–al* (114/1.451), como se aprecia en los ejemplos 33 y 34 de ambos corpus, aunque el primer sufijo es más frecuente en el corpus de análisis y el segundo en el corpus lexicográfico.

33. cervicogénico, ectópico, farmacológico, galvánico, hematológico, idiopático, laberíntico, miocárdico, quístico, urémico, vitamínico, ótico, úrico.
34. bulboespinal, duodenal, ecuatorial, facial, helicoidal, monoclonal, neurosensorial, oculofacial, sinovial, tubulointersticial, uretral, yeyunal.

Salvo por los sufijos *–ino* y *-udo*, los 10 primeros sufijos del corpus de análisis son los mismos 10 primeros sufijos del corpus lexicográfico si bien el orden no es el mismo. Estos sufijos son en orden de frecuencia: *-ico*, *-al* (*-ar*), -

[57] El estudio de Gallegos tiene un corpus relativamente pequeño ya que se compone de 2 textos completos y 3 capítulos del texto científico (2003: 42). Por eso, no es posible afirmar que sus hallazgos puedan generalizarse en el conjunto de la lengua o incluso en los ámbitos de especialidad.

nte, *-ble*, *-eo*, *-ario*, *-ino, -ivo, -oso* y *-udo*. Semánticamente, todos estos son sufijos adjetivales que indican relación con, o cualidades y propiedad de personas, animales o cosas (Varela 2005: 55), como se observa en los ejemplos 35 a 42.

35. creciente, insulinodependiente, dominante, recidivante, necrosante
36. computable, cotizable, desgravable, programable, sostenible, susceptible
37. aéreo, cutáneo, eutiroideo, faríngeo, laríngeo, percutáneo, raquídeo
38. arancelario, dentario, leucocitario, mamario, portuario, urinario
39. exocrino, femenino, intrauterino, murino, uterino, equino, tromboplastino
40. activo, auditivo, conjuntivo, digestivo, radioactivo, recesivo, depresivo
41. canceroso, cartilaginoso, escamoso, fibroso, racemoso, ulceroso, venoso
42. agudo, cabelludo, desnudo, estornudo, subagudo

Según Varela (2005: 50) los adjetivos también se pueden clasificar según la categoría gramatical de base. Así, los adjetivos más productivos de este estudio pueden clasificarse básicamente en adjetivos deverbales en orden de frecuencia (*-nte, -ble, -ivo*) como los casos del ejemplo 43 y adjetivos denominales (*-al* (*-ar*), *-ario*, *-ico, -ino, -ivo*[58], *-oso, -udo*) como los casos de ejemplo 44.

43. dominante, codificante, palpable, absorbible, adhesivo, agresivo
44. bulboespinal, mitocondrial, capilar, molecular, pigmentario, placentario, seborreico, pancreático, uterino, cristalino, adhesivo, cohesivo, calloso, cerebeloso, cabelludo, desnudo.

Obsérvese que no se han incluido los adjetivos derivados del sufijo *–ado*, ya que se ha querido hacer la diferencia entre los adjetivos generales y los derivados como participios de verbos que se forman a partir de oraciones pasivas, que se analizan a continuación.

[58] Este sufijo también puede formar adjetivos de base verbal como en los casos de nutritivo, competitivo, etc.

A continuación, están los participios con aproximadamente un 4,98% en el corpus de análisis y un 0,8% en el corpus lexicográfico. Esta diferencia se explica porque en el corpus lexicográfico no se diferenció los participios de pasado de los adjetivos terminados en *–ado* ya que el etiquetador que se empleó no hizo tal distinción en los diccionarios. Aún así la presencia de adjetivos/participios derivados de verbos es muy importante en el corpus de diccionarios, como se ilustra en los ejemplos de 45 con 123 casos, lo que lo situaría entre los 10 sufijos más comunes si se tiene en cuenta como adjetivo.

45. yodado, acumulado, administrado, almacenado, indiferenciado, insaturado, integrado, cultivado, amortizado, automatizado, cayado, contorneado, controlado, etc.

Sin embargo, debido a su procedencia oracional, las palabras terminadas en *–ado* se han tenido en cuenta en este estudio como participios. Estos participios forman estructuras sintagmáticas que son fruto de una oración relativa en voz pasiva, como ocurre en los ejemplos 46 y 47.

46. agua destilada estéril (agua que se destila de modo estéril)
47. gen mitocondrial relocalizado (gen mitocondrial que se ha relocalizado)

Como sucede en inglés, la mayoría de adverbios se deriva o se comporta como adjuntos de adjetivos, como se observa en los ejemplos de 48 y 49.

48. linfocitos inmunológicamente activos (inmunológicamente> inmunológico)
49. cáncer vesical cistoscópicamente visible (cistoscópicamente> cistoscópico)

Morfológicamente, la mayoría de adverbios terminan en *–mente,* como en los ejemplos de 50.

50. absolutamente, fuertemente, generalmente, genéticamente, gravemente, inmediatamente, inmunológicamente, irrestrictamente, mayoritariamente, serológicamente, sexualmente, temporariamente, totalmente.

De acuerdo con la tipología de Kaul (2002: 144), hay 14 casos de adverbios que son de punto de vista, en los que señala el campo de referencia de la propiedad modificada, como se ejemplifica en 51.

51. cistoscópicamente, culturalmente, genéticamente, evolutivamente, asintóticamente, inmunológicamente, serológicamente, químicamente, sexualmente, térmicamente, triangularmente, consensualmente, automáticamente.

También pueden destacarse los adverbios de intensificación de grado (10 casos) y aspectuales (4 casos), como se ilustra en el ejemplo 52.

52. absolutamente, altamente, fuertemente, gravemente, puramente, rápidamente, completamente, totalmente, parcialmente, mayoritariamente.

Desde un punto de vista sintáctico, los adverbios terminados en *-mente* funcionan sintácticamente como modificadores directos de adjetivos, participios de pasado o adverbios, como se observa en los ejemplos 53 y 54.

53. confirmación de azoospermia en varones **sexualmente maduros** (a un adjetivo).
54. economías asiáticas **recientemente industrializadas** (a un participio)[59].

Estos sintagmas con adverbios y adjetivos o participio de pasado generalmente se originan a partir de oraciones del tipo relativa pasiva como en el ejemplo 55.

55. economías asiáticas recientemente industrializadas (economías asiáticas que se han industrializado recientemente).

[59] Ejemplo tomado del diccionario Routledge de economía y finanzas.

Finalmente, puede concluirse que la tendencia nominalizadora del discurso científico-técnico no sólo se observa en la gran cantidad de sustantivos en la modificación, sino también en la procedencia nominal de muchos adjetivos y adverbios, como se ha presentado en los datos.

Cabe destacar la presencia de preposiciones dentro de las categorías cerradas. La preposición *de* predomina ampliamente en ambos corpus con un 78,65% y 76,9% en el corpus de análisis y lexicográfico respectivamente. Luego sigue la preposición *en* con un 7, 06% y 5,86%, respectivamente. La preposición *con* representa el 6,31 en el corpus de análisis y la preposición *a* con un 5,33. Todas las otras preposiciones del corpus de análisis (por, a, para, mediante, sin, sobre, contra, desde, tras, entre) y del corpus lexicográfico (por, con, para, sobre, sin, contra, bajo, según, mediante, vía, durante, desde, tras) no superan el 8 y 12%, respectivamente.

Este trabajo refrenda hasta cierto punto lo que Estopà (1999: 100) ha encontrado para el catalán sobre el uso de las preposiciones en las unidades terminológicas. Aunque este trabajo ha tomado un tipo de unidad de análisis menos restringida que el trabajo de Estopà, los datos sobre el uso de las preposiciones coinciden:

> "Vam remarcar que la preposició de és la més usada en els discursos especialitzats, molt més que en altres llengües romàniques com el francès en què el camp semàntic de la preposició de es reparteix formalment entre la preposició de i la preposició à [L'Homme 1996b]."

Si bien en esta tesis se reporta un uso importante de otras preposiciones diferentes de *de* no sólo en el corpus (21,36%) sino en los diccionarios (23,1%).

5.5.3 Frecuencia de los patrones por aparición

En este apartado, se comparan los 20 patrones más frecuentes de ambos corpus. Puede observarse en la tabla 20 que en ambos corpus el patrón más frecuente es el patrón N Prep N Adj, pero con diferente porcentaje de ocurrencias (31,66% y 15,17%), observándose un predominio en el corpus de análisis. Sin embargo, se explica que el porcentaje sea mucho menor en el corpus lexicográfico ya que la cantidad de patrones es también mucho mayor. Y, por tanto, la distribución es más equitativa. Después aparecen los tres siguientes patrones más frecuentes en el corpus de análisis (N Adj Prep N, N Prep N Prep N y N Adj Adj) son, de cierto modo, los más frecuentes en el corpus lexicográfico, pero no en el mismo orden (quinto, tercero y segundo, respectivamente). Así el patrón Prep N Prep N es el tercero más frecuente en ambos corpus con porcentajes muy similares (10,68% y 10,37%). Estos patrones representan de la mayoría de datos en ambos corpus (84,34% y 66,95%) aunque mucha más en el corpus de análisis que en el lexicográfico. En este último, los patrones están más distribuidos mientras que en el de referencia hay diferencias importantes en algunos casos.

Posteriormente, el orden en ambos corpus comienza a variar, pero en los primeros 10 patrones hay 4 (N Prep N Adj, N Adj Prep N, N Prep N Prep N, N Adj Adj) patrones iguales y 6 diferentes. En los segundos 10 patrones hay 4 patrones iguales (N Prep N Prep N Prep N, N Prep N Adj Adj, N Adv Adj, N Adj N) y el resto no coinciden independientemente del orden. Hay 3 patrones que aparecen en una u otra franja con los cual hay 11 patrones comunes entre ambos corpus.

Los patrones de la segunda franja representan el 8,49% y 12,48% de todos los patrones de manera distribuida en ambos corpus. En total, los 20 primeros patrones de ambos corpus representan el 92,83% y 79,43% de los sintagmas de ambos corpus, con lo cual puede considerarse que los sintagmas que representan estos patrones son candidatos potenciales a lexicalizarse.

Tokens	Patrón corpus	Porcentaje	Tokens	Patrón Dic.	Porcentaje
3	N Prep N Adj	31,66	3	N Prep N Adj	15,17
3	N Adj Prep N	16,13	3	N Adj Adj	10,64
3	N Prep N Prep N	10,68	3	N Prep N Prep N	10,37
3	N Adj Adj	6,73	3	N Prep N N	7,76
3	N Adj PP	4,88	3	N Adj Prep N	7,67
4	N Adj Prep N Adj	3,68	4	N Prep N Prep N Adj	3,76
4	N Prep N Adj Prep N	3,31	3	N N Adj	3,28
3	Adj N Prep N	2,76	4	N Adj Prep N Adj	3,09
4	N Adj Prep N Prep N	2,67	3	N N N	2,67
3	N Prep Adj N	1,84	4	N Adj Prep N Prep N	2,54
3	N PP Prep N	1,29	4	N Prep N Prep N Prep N	2,54
4	N Prep N Prep N Prep N	1,11	4	N Prep N Adj Prep N	1,66
4	N Prep N Prep N Adj	1,01	4	N Prep N Adj Adj	1,51
3	N Prep N PP	0,93	3	N N Prep N	1,37
4	N Prep N Adj Adj	0,83	3	N Adj N	1,21
4	N Adj PP Prep N	0,74	3	N Prep Adj N	1,15
3	N N Adj	0,74	3	N Adv Adj	1
5	N Adj Prep N Prep N Adj	0,64	4	N Prep N Adv Adj	0,74
3	N Adv Adj	0,65	4	N Adj Adj Adj	0,73
3	N Adj N	0,55	4	N Prep N Prep Adj N	0,57

Tabla 20: Comparación de los primeros 20 patrones del corpus y el corpus lexicográfico en español.

Desde un punto de vista terminológico y traductivo, los 20 primeros patrones de ambos corpus son de extremada relevancia por las siguientes razones. En primer lugar, son estos patrones los que generarían más candidatos a término. En segundo lugar, son los que tenderían a aparecer más en los repositorios terminológicos. En tercer lugar, serían los tipos de sintagmas que más aparecerían en las traducciones y por ende darían más problemas al traductor. Cuarto, su sistematización sería de gran ayuda no sólo al terminólogo o lexicógrafo especializado sino al traductor. Finalmente, sería un factor más de ponderación para la extracción de términos, recuperación de información y traducción automática, entre otros.

Todos los patrones del corpus de análisis aparecen en los patrones de los diccionarios, pero no al contrario debido a que se sacó una muestra de estos al ser un corpus de análisis.

5.5.4 Frecuencia de los patrones por longitud

En cuanto a su longitud, los patrones de 3 tokens predominan sobre los de 4 en la primera franja de 10 patrones en ambos corpus (7 y 9, respectivamente). En la segunda franja de 10 patrones, el predominio de los patrones de 3 tokens es menos evidente ya que hay 5 patrones de 3 tokens en el corpus de análisis y 6 en el corpus lexicográfico contra 4 patrones de 4 tokens en ambos corpus. Sólo aparece un patrón de 5 tokens en ambos corpus (N Adj Prep N Prep N Adj).

Así puede decirse que predominan los patrones de 3 tokens con 12 y 11 patrones en cada corpus. Estos patrones de 3 tokens representan 78,84% y 62,29%, respectivamente. Después siguen los patrones de 4 tokens con 7 y 9 patrones y representan un 13,35% y 17,14%, respectivamente. Esto, muestra una vez más que la extensión está ligada a la frecuencia, como se ha explicado anteriormente.

5.6 Contraste de los resultados con los patrones encontrados con los del Crea de la RAE

En este último apartado se contrastan los resultados en español con los datos del Corpus del Español Actual (CREA) de la Real Academia Española. Para ello, se ha solicitado una consulta con los mismos patrones que se usaron en la primera extracción de datos. Dicha consulta se realizó sobre un corpus de 5.397 documentos y 143.440.437 tokens.

En la tabla 21, puede observarse que de los 20 patrones más frecuentes del corpus CREA, los patrones de 3 tokens son los más frecuentes con 12

patrones que representan el 78,22% y luego los patrones de 4 tokens con 16,48%.

Entre los 3 corpus, existen 12 patrones comunes (N Adj Prep N, N Prep N Adj, N Prep N Prep N, N Adj Adj, N Prep Adj N, N Adj Prep N Adj, N Prep N Adj Prep N, N Adj N, N Adj Prep N Prep N, N Prep N Prep N Prep N, N N Adj y N Adv Adj). Estos patrones representan en cada corpus más de la mitad de todas las ocurrencias (74,12% en el corpus de análisis, 51,67% en el corpus lexicográfico y 68,92% en el corpus CREA).

Tokens	Patrón corpus	Porc.	Tokens	Patrón Dic.	Porc.	Tokens	Patrón Crea	Porc.
3	N Prep N Adj	31,66	3	N Prep N Adj	15,17	3	N Adj Prep N	17,4
3	N Adj Prep N	16,13	3	N Adj Adj	10,64	3	N Prep N Adj	14,4
3	N Prep N Prep N	10,68	3	N Prep N Prep N	10,37	3	N Prep N Prep N	13,3
3	N Adj Adj	6,73	3	N Prep N N	7,76	3	N Prep N PP	10,4
3	N Adj PP	4,88	3	N Adj Prep N	7,67	3	Adj N Prep N	6,97
4	N Adj Prep N Adj	3,68	4	N Prep N Prep N Adj	3,76	3	N Adj Adj	5
4	N Prep N Adj Prep N	3,31	3	N N Adj	3,28	4	N Prep Adj N	3,9
3	Adj N Prep N	2,76	4	N Adj Prep N Adj	3,09	4	N Adj Prep N Adj	3,56
4	N Adj Prep N Prep N	2,67	3	N N N	2,67	3	N Prep N Adj Prep N	2,97
3	N Prep Adj N	1,84	4	N Adj Prep N Prep N	2,54	4	N Adj N	2,23
3	N PP Prep N	1,29	4	N Prep N Prep N Prep N	2,54	3	N Adj Prep N Prep N	2,22
4	N Prep N Prep N Prep N	1,11	4	N Prep N Adj Prep N	1,66	3	N Adj PP	1,95
4	N Prep N Prep N Adj	1,01	4	N Prep N Adj Adj	1,51	4	N N Prep N	1,95
3	N Prep N PP	0,93	3	N N Prep N	1,37	3	N Prep N Prep N Prep N	1,54
4	N Prep N Adj Adj	0,83	3	N Adj N	1,21	4	N N N	1,48
4	N Adj PP Prep N	0,74	3	N Prep Adj N	1,15	4	N N Adj	1,42
3	N N Adj	0,74	3	N Adv Adj	1	3	N Prep N Prep N Adj	1,23
5	N Adj Prep N Prep N Adj	0,64	4	N Prep N Adv Adj	0,74	4	N Adj Prep Adj N	1
3	N Adv Adj	0,65	4	N Adj Adj Adj	0,73	4	N Adv Adj	0,94
3	N Adj N	0,55	4	N Prep N Prep Adj N	0,57	3	N Adj PP Prep N	0,76

Tabla 21: Comparación de los primeros 20 patrones del corpus de análisis, el corpus lexicográfico y el CREA.

Si se compara el corpus de análisis con el corpus CREA se puede observar que los tres primeros patrones del corpus de análisis (N Prep N Adj, N Adj Prep N, N Prep N Prep N) son los tres patrones más frecuentes del corpus CREA, pero varía el orden en los dos primeros. Estos tres patrones representan de más de la mitad de las ocurrencias en el corpus de análisis (58,47%) y casi la mitad de las ocurrencias en el corpus CREA (48,25%). Sin embargo, el porcentaje del

patrón más frecuente del corpus de análisis (N Prep N Adj) es mucho más frecuente que el del corpus CREA (31,66% vs. 15,4%). Además, la diferencia entre el primero y segundo patrón de cada corpus es muy considerable (15,53%) en el corpus de análisis, mientras en el corpus CREA, dicha diferencia es poca (3,2%), lo que muestra el predominio de dicho patrón en el corpus de análisis.

A pesar del predominio de este patrón, no sólo estos tres patrones sino también los 12 que comparten los tres corpus y los 20 más frecuentes de cada corpus son estructuras de la lengua general y no son exclusivas de los ámbitos de especialidad no sólo en el corpus sino en el corpus de diccionarios. Por tanto, la gramática de los llamados "lenguajes de especialidad" debe explicarse perfectamente desde la gramática de la lengua general.

5.7 Recapitulación

En este capítulo, se han discutido los resultados del análisis del corpus de análisis en español y el contraste con el corpus lexicográfico y el corpus CREA de la RAE.

1. En cuanto a la longitud de los sintagmas, el corpus de análisis en español, al igual que el corpus del inglés, los patrones de 3 tokens predominan ampliamente con un 80,66% de todas las ocurrencias, seguido de los patrones de 4 tokens con 16,74%. Los patrones de 5 y 6 tokens tan solo representan a un 2,586% de las ocurrencias.

2. En cuanto a la categoría gramatical modificadora que predomina en español es igualmente el sustantivo con un 30,37%, seguido por el adjetivo con un 26,81%. Es importante resaltar la presencia de otras categorías léxicas como los participios con 2,56% y los adverbios con 0,32%. En el caso de las categorías cerradas, las preposiciones representan el 29,47% de la modificación del sintagma. Dadas las características sintácticas del español, el uso frecuente del sustantivo como modificador se justifica plenamente ya que hacen parte de los sintagmas preposicionales que modifican al núcleo y son la forma natural de expandir un sintagma en español.

3. En cuanto a los patrones más frecuentes, los patrones más frecuentes son N Prep N Adj con un 31,66%, N Adj Prep N con un 16,13%, N Prep N Prep N con un 10,68%, el patrón N Adj Adj con 6,73% y el patrón N Adj PP con 4,88%. Estos cinco patrones representan el 70,08% de todas las ocurrencias del corpus y por tanto, presentan menos variación sintáctica. Entre los patrones de 4 tokens cabe destacar los patrones N Adj Prep N Adj con 3,68% y N Prep N Adj Prep N con 3,31%. Estos patrones presentan más variación y representan a muchas menos ocurrencias dentro del corpus.

4. De acuerdo con la dependencia sintáctica, la relación de dependencia [A [B C]] es la más frecuente de la muestra en español como más del 50,5% de todas las ocurrencias (101) del corpus de análisis, seguida de la relación de dependencia [[A B] C] con un 45,5% de todas las ocurrencias (91). Por último, la dependencia [[A B] [C D]] representa el 3% de todas las ocurrencias (6) para patrones de 4 tokens. Los patrones que presentan una única relación de dependencia son: N Prep Adj N, N Adj Prep N, N Adj PP, N Adj Adj y Adj N Prep N y N Adj Prep N Adj. De estos, N Adj Prep N, N Adj PP, N Adj Adj y Adj N Prep N tiene la misma forma de dependencia sintáctica [[A B] C]. El único patrón que tiene la dependencia sintáctica [A [[B C]] es N Prep Adj N. El patrón de 4 tokens, N Adj Prep N Adj tiene la dependencia [[A B] [C D]]. Los patrones que tienen dos relaciones de dependencia sintáctica son: N Prep N Prep N y N Prep N Adj. En el patrón N Prep N Prep N, la relación de dependencia [A [[B C]] representa 70,37% de las ocurrencias y [[A B] C] al 29,62%. En el caso del patrón N Prep N Adj, la dependencia [A [[B C]] representa el 89,53% ocurrencias y [[A B] C] al 9,30%).

5. El contraste con el corpus lexicográfico y el CREA de la RAE corrobora los resultados obtenidos en el corpus de análisis en cuanto a la longitud y frecuencia de los sintagmas, predominio de patrones y categoría léxica en la premodificación. Estos resultados no son exclusivos de los ámbitos de especialidad sino que se inscriben dentro de la gramática de la lengua general. Igualmente, los resultados de los diccionarios permiten afirmar que los análisis hechos se pueden extrapolar a otras áreas del conocimiento y que no son exclusivos de las ciencias “duras”.

Al igual que en inglés, se ha demostrado que la existencia de los SNEE es una característica de la lengua que puede presentarse con mayor frecuencia en el discurso especializado y que, además, pueden describirse, clasificarse, explicarse y predecirse desde la gramática de una lengua como todos los fenómenos lingüísticos de los discursos de los ámbitos de especialidad, como lo plantea la teoría comunicativa de la terminología.

6. Descripción y análisis semántico de los patrones en inglés

6.1 Introducción

En este capítulo se pretende caracterizar semánticamente los sintagmas nominales especializados extensos en inglés de modo que conjuntamente con los resultados sintácticos permitan interpretar estos sintagmas usando elementos lingüísticos obtenidos de las regularidades observadas.

Aunque no se pretende establecer las relaciones semánticas entre los diferentes constituyentes del sintagma a partir de un conjunto cerrado de clases y relaciones semánticas como lo ha hecho Oster (2005), se establecerán patrones semánticos de los sintagmas empleando WordNet 2.1 y UMLS. Estos patrones permitirán observar las tendencias semánticas en inglés al menos para este campo temático.

Se empleó WordNet 2.1 y UMLS para etiquetar el corpus por varias razones. En primer lugar, WordNet 2.1 es la ontología más empleada en todos los estudios relacionados con la lingüística y en la construcción de otros recursos lingüísticos como diccionarios, tesauros, etc. y el procesamiento del lenguaje natural. Por otro lado, UMLS es el recurso más empleado en las ciencias de la salud para la indexación de bibliotecas, procesamiento del lenguaje biomédico y estudios lingüísticos relacionados con la medicina. En segundo lugar, el acceso a ambas herramientas es de acceso libre bajo licencia respectiva. En tercer lugar, el uso de ambas ontologías para el inglés permitía obtener más generalizaciones puesto que WordNet 2.1 es una ontología para propósitos generales con una buena cantidad de entradas en medicina pero UMLS es una ontología especializada en ciencias de la salud.

6.2 Criterios y selección del corpus de análisis en inglés

Como se explicó en el capítulo de la metodología (§3.7), para el análisis semántico se seleccionó manualmente una muestra a partir de la muestra del análisis morfosintáctico (1.080) que a su vez es la misma muestra empleada para el análisis de dependencias en inglés de §4.3.5. Para ello, se seleccionó un 24,37%[60] de los sintagmas que equivalen al 88,82% de todas las ocurrencias del corpus de análisis y se distribuyó proporcionalmente de acuerdo con su frecuencia como se hizo con la muestra sintáctica. Es decir, al patrón más frecuente, le corresponden más sintagmas para el análisis semántico y al patrón menos frecuente se le asignan menos sintagmas. Por ejemplo, el patrón Adj N N es el más frecuente del corpus y le corresponden 54 sintagmas y el menos frecuente es el N Adj N y le corresponden 5 sintagmas de la muestra.

	Muestra sintáctica	Muestra semántica	Porcentaje
Inglés	1.096	232 SN	24,37%

Tabla 1: Muestra de sintagmas extraídos para el análisis semántico.

Puesto que para el análisis sintáctico se había seleccionado los sintagmas de mayor a menor frecuencia hasta completar la muestra que correspondía a cada patrón sintáctico, los sintagmas para la muestra semántica se seleccionaron siguiendo el mismo criterio de distribución de la muestra.

Debido a que la forma de completar la muestra se basa en la frecuencia y la distribución proporcional entre patrones, cada muestra está ajustada y no se

[60] Aunque una muestra de 232 sintagmas puede parecer limitada, la decisión del 24,37% se basa en la selección de un 20% de la muestra de forma completa para cada patrón y por eso se ha llegado hasta ese porcentaje. Aunque está limitación se justifica por la dificultad de etiquetar a mano y en dos sistemas diferentes los tokens de estos sintagmas (un total de 709 tokens en cada sistema). Además, cada token se ha buscado, desambiguado en otros diccionarios, y observado en contexto para poder asignar la etiqueta de WordNet o UMLS de manera correcta. Para más información, véase §3.4 y §6.3 de este capítulo.

limita al porcentaje exacto que se ha calculado (e.g. no se puede seleccionar media ocurrencia u ocurrencia y media para determinado patrón).

6.3 Metodología

La muestra seleccionada se etiquetó manualmente para un total de 3.323 tokens. Las categorías semánticas, presentadas en §3.2.2.2 y §3.2.2.4, se asignaron por separado en todos los núcleos, en primer lugar, con WordNet 2.1 (*synsets*) y en segundo lugar, con UMLS 2006AB (*semantic types* y *semantic groups*). Posteriormente, se etiquetó la premodificación de acuerdo con las categorías léxicas: sustantivos, adjetivos y adverbios, de modo que pudiera asegurarse la consistencia de etiquetaje dentro de cada categoría léxica y clase semántica tanto en WordNet 2.1 como en UMLS 2006AB. Por ejemplo, el sustantivo *chromosome* comparte la misma clase semántica (*noun.body*) con el adjetivo *chromosomal*.

Si se encontraba más de una posibilidad de etiquetaje se asignaba la clase que correspondiera más al significado del sintagma o en caso de ambigüedad se dejaban las diferentes clases como en el caso de *factor* que tiene las clases *noun.body/noun.cognition/noun.event* de WordNet. En cualquier caso, la primera opción del etiquetaje se dejaba como la más próxima al significado del sintagma. En el caso de UMLS, se presenta más de una opción, las cuales se marcan como primera y segunda opciones en el sistema. Generalmente, ambas opciones están relacionadas semánticamente, con lo cual no hubo problemas de ambigüedad. Por ejemplo, *kinase* pertenece a las clases *Amino Acid, Peptide, or Protein*, *Enzyme* y *Substance* que están relacionadas jerárquicamente y pertenecen al mismo tipo semántico, *Chem* (*Chemicals&Drugs*).

Luego se tabularon los datos en Statgraphics 5.1 en cuanto a las clases semánticas[61] de los núcleos, la premodificación tanto en WordNet 2.1 y posteriormente con UMLS. De allí, se seleccionaron las categorías más prototípicas de la muestra tanto para los núcleos como para la premodificación.

Se obtuvo luego el conjunto de patrones semánticos más frecuentes y se compararon con los patrones sintácticos, de modo que pueda verse la relación entre los aspectos sintáctico y semántico y las tendencias que presentaban.

Igualmente se analizó la frecuencia y las tendencias en cuanto a la posición dentro de la premodificación y de acuerdo con la categoría léxica (N, Adj y PP).

6.4 Resultados

6.4.1 Análisis de las clases semánticas de los núcleos según WordNet 2.1

Como puede verse en la tabla 2, la clase semántica más frecuente en el corpus de análisis es *body* con un 18,53%, seguido de *substance* con un 15,95% y *act* con un 10,34%. Estas tres clases semánticas representan el 44,82% de todos los núcleos de la muestra.

Como puede verse en el ejemplo 1, y como lo indica el nombre de la clase, todos se refieren a partes del cuerpo.

1. gene, allele, area, blood, brain, chromosome, collagen, cortex, follicle, hormone, membrane, myocardium, region, serum, sheath, system.

[61] Se empleará el sintagma "clase semántica" para referirse a los "synsets" de WordNet y "semantic types" de UMLS.

Sin embargo, obsérvese que también existen algunas sustancias como *serum* y *hormone* que podrían estar bajo la clase *substance* como se ve en las definiciones del diccionario Steadman.

> serum: a clear watery fluid, especially that moistening the surface of serous membranes, or exuded in inflammation of any of those membranes.
>
> hormone: a chemical substance, formed in one organ or part of the body and carried in the blood to another organ or part; depending on the specificity of their effects, hormone's can alter the functional activity, and sometimes the structure, of just one organ or of various numbers of them.

Igualmente, existe algunas clases que podrían considerarse como hiperónimos ya que no son exactamente partes del cuerpo sino tipos de paratérminos que se refieren a lugares más amplios no definidos que podrían albergar a partes del cuerpo como es el caso de *area*, *region* y *system*.

En el ejemplo 2, se presentan casos de la clase *substance* del cuerpo o de seres vivos.

2. DNA, RNA, albumin, buffer, enzyme, fiber, kinase, mRNA, molecule, oligosaccharide, platelet, pol, polymerase, product, protein, residue, sulphate.

En el caso de los ejemplos *residue* y *product* no se refieren propiamente a sustancias sino a materias que funcionan como hiperónimos como reza la definición del diccionario Steadman.

> product: anything produced or made, either naturally or artificially.
>
> residue: that which remains after removal of one or more substances.

Se deduce entonces que en WordNet existen inconsistencias en el tratamiento de algunos casos como las sustancias ya que *enzyme* se considera una sustancia pero *hormone* una parte del cuerpo.

En 3, se presentan los ejemplos de los sustantivos de la clase *act* (de acción). Obsérvese que todos los sustantivos son deverbales (nominalizaciones por derivación) o son producto de una conversión categorial de un verbo a sustantivo (e.g., *transfer, test*).

3. breakage, contraction, hybridization, infusion, ligation, medication, process, production, response, spectrometry[62], test, transfer, transition.

Clase semántica	Frecuencia	Porcentaje
noun.body	43	18,53
noun.substance	37	15,95
noun.act	25	10,77
noun.group	17	7,33
noun.process	17	7,33
noun.attribute	16	6,9
noun.cell	12	5,17
noun.state	12	5,17
noun.artifact	10	4,31
noun.animal	8	3,45
noun.object	7	3,02
noun.cognition	6+1	2,59
noun.event	6	2,59
noun.location	5	2,16
noun.phenomenon	5	2,16
noun.quantity	3	1,29
noun.food	1	0,43
noun.person	1	0,43
noun.relation	1	0,43
noun.shape	1	0,43

Tabla 2: Clases semánticas de los núcleos en WordNet 2.1.

También es importante destacar las clases *group* (7,33%), *process* (7,33%) y *attribute* (6,9%), como se ejemplifica en 4, 5 y 6.

4. collection, clone, cluster, library, line, sequence, strain.

[62] En el caso del compuesto culto *spectrometry*, se considera un préstamo del alemán de acuerdo con el diccionario Webster (Ger spektrometer: see SPECTRO- & -METER).

5. activation, activity, association, chromatography, digestion, electrophoresis, expression, growth, immunoreactivity, metabolism, purification, reaction, replacement, variation.

6. analogues, difference, distance, distribution, expression, fidelity, identity, level, mass, phenotype, responsiveness, size, weight.

Los ejemplos de 4 son todos casos de sustantivos colectivos y los ejemplos de 5 son sustantivos deverbales como sucede con la clase *act*, lo cual se puede evidenciar en los tipos de sufijos: *-ation*, *-ity* e *-ion*. Los ejemplos de la clase *attribute* se refieren a sustantivos que denotan cualidad de un objeto como sucede con *size*, *weight*, etc.

Es importante anotar que las clases *act* y *process* no son las más frecuentes en la premodificación pero si en los núcleos lo que demuestra que estas nominalizaciones tienden a ir en el núcleo y los atributos (*attribute*, 9,01%) en la premodificación así como los estados (*state*, 4,4%) y adjetivos generales (*adj.all*, 3,35%).

Finalmente, en 7 se recogen ejemplos de aquellas unidades léxicas (N, Adj y Adv) que no son términos pero que pueden ayudar en la identificación de unidades terminológicas o que adquieren carácter terminológico dentro de un contexto o área determinados y que pueden denominarse paratérminos.

7. area, factor, product, region, system, type, central, complex, critical, dependent, dominant, double, effective, false, heavy, high, large, modern, negative, normal, partial, specific, total, slightly, right, highly

6.4.2 Análisis de las clases semánticas de la premodificación según WordNet 2.1

En la tabla 3, se presentan las clases semánticas de la premodificación en WordNet 2.1.

La clase semántica más frecuente es *substance* con un 21,59% de todas las ocurrencias, seguida por la clase *body* con un 15,72% y luego la clase *animal* con un 11,53% del total de las ocurrencias. En conjunto, representan un 48,84% de toda la premodificación.

Clase semántica	Frecuencia	Porcentaje
noun.substance	103	21,59
noun.body	75	15,72
noun.animal	55	11,53
noun.attribute	43	9,01
noun.state	21	4,4
noun.cell	18	3,77
adj.all	16	3,35
noun.artifact	15	3,14
noun.group	13	2,73
noun.process	12	2,52
noun.act	11	2,31
verb.change	10	2,1
noun.object	9	1,89
noun.cognition	7	1,47
noun.event	7	1,47
noun.location	6	1,26
noun.phenomenon	5	1,05
verb.cognition	5	1,05
noun.person	3	0,63
verb.communication	3	0,63
noun.quantity	2	0,42
noun.shape	2	0,42
verb.contact	2	0,42
adv.all	1	0,21
noun.food	1	0,21
noun.plant	1	0,21
noun.relation	1	0,21
verb.perception	1	0,21
verb.motion	1	0,21
notWN	28	5,87

Tabla 3: Clases semánticas de la premodificación en WordNet 2.1.

En 8, 9 y 10, se listan los ejemplos de estas tres clases.

8. cytokine, DNA, glucose, H1, HLA-DR2, MRP, MUC7, NGF, protein, amino acid, albumin, amyloid, androgen, cDNA, calcium, carbon, glucose, kinase, myosin, nucleotide, phosphatase, phosphate, platelet, polymerase, polypeptide, potassium, promoter, sodium, tau, tetrachloride, transcript, tyrosine.

9. allele, backbone, bile, blood, bone, brain, channel, chromosome, complex, collagen, corpuscle, duct, gene, insulin, kidney, lung, marrow, membrane, muscle, plasma, receptor, root, serum, T.

10. bovine, calf, embryonic, eukaryotic, fetal, HBV, herpes, HIV-1, human, hybrid, male, mouse, mutant, ovine, rat.

Al igual que en los núcleos, existen las mismas inconsistencias con las clases *substance* y *body* ya que *bile*, *blood*, *insulin*, *plasma* y *collagen* pueden pertenecer a la clase *substance* y no a la clase *body*, como se deduce de las definiciones del diccionario LDOCE[63].

bile: a bitter green-brown liquid formed in the liver, which helps you to digest fats.

blood: the red liquid that your heart pumps around your body.

insulin: a substance produced naturally by your body which allows sugar to be used for energy.

collagen: a protein found in people and animals.

plasma: 1 the yellowish liquid part of blood that contains the blood cells. 2 technical the living substance inside a cell.

[63] Longman Dictionary of Contemporary English en CD-ROM, 2003.

En el caso de la clase *animal*, los ejemplos son ilustrativos y se refieren a virus, mamíferos o a partes de seres vivos.

Obsérvese que las clases *substance* y *body* son igualmente las más frecuentes no sólo en la premodificación sino en los núcleos, salvo por la variación en el orden. Esto muestra una fuerte afinidad entre la semántica de los sintagmas y el ámbito temático en que se inscriben.

Igualmente, cabe destacar también las clases *attribute* con un 9,01%, *state* con un 4,4% y *cell* con un 3,77% y sus respectivos ejemplos en 11, 12 y 13.

11. adipose, artificial, central, contiguous, control, dense, effective, green, heavy, high, highly, individual, large, length, linkage, mass, modern, multiple, natural, permeability, red, sexual, smooth, specific, trait, white.

12. consensus, critical, dependent, disease, false, hepatitis, hepatoma, immature, maternal, medial, normal, partial, paternal, resistance, TLE, UPD.

13. adipocyte, cell, cellular, leukocytic.

En el caso de la clase *attribute*, todos los ejemplos corresponden a sustantivos y adjetivos que denotan principalmente atributos humanos y de objetos.

La clase *state* denota enfermedades (*hepatitis*, *hepatoma*) o estados de enfermedades (*critical, dependent, normal, partial,* etc.) o estados estables (*paternal, maternal*).

En cuanto a la clase *cell*, todos los ejemplos se refieren a células o partes de ellas. En este sentido, adjetivos como *eukaryotic* etiquetados como animal deberían estar relacionados a la clase *cell*, así como los ejemplos *gene*, *allele* y *chromosome* etiquetados bajo la clase *body* como se deduce de las definiciones de los diccionarios LDOCE y Steadman.

gene: a part of a cell in a living thing that controls what it looks like, how it grows, and how it develops.

allele: any one of a series of two or more different genes that may occupy the same locus on a specific chromosome.

eukaryotic: pertaining to or characteristic of a eukaryote. A cell containing a membrane-bound nucleus with chromosomes of DNA, RNA, and proteins [...].

Es importante resaltar que no se le asignó ninguna clase semántica en WordNet 2.1 a más del 5,37% de los casos y a 63 registros de 3.089 (2,03%) no se le encontró la clase semántica directamente pero se le asignó con la ayuda de los diccionarios de referencia de medicina. Así, el total de ocurrencias que no se encuentran en WordNet 2.1 se eleva a un 7,4%, lo que puede afectar el etiquetaje automático substancialmente. Por eso, se decidió etiquetar manualmente para reducir estos silencios de acuerdo con WordNet 2.1 pero señalando dichas unidades como no encontradas (*not found*) o NotWN. Muchas de estas unidades léxicas son siglas y términos muy especializados que no se encuentran en muchos casos ni en diccionarios generales ni especializados, como se observa en los ejemplos de 14.

14. ArG, BD, biotinylated, CaCo2, electrospray, ELT-3, etl2, Genius, helper, HPS, IGF2R, IL11RA, immunoreactive, K2, Kozak, MHC, MUL, multipoint, PAR, PCR, pseudocontact, SDS, TA, TATAA, TnTf, TSC2-null, X.

6.4.2.1 Clases semánticas de acuerdo con la posición dentro de la premodificación

Si se observan las clases semánticas de la premodificación de acuerdo con la posición dentro del sintagma se deduce que no existen diferencias importantes ya que tienden a predominar las mismas clases semánticas. Esto explica que las clases semánticas encontradas están estrechamente relacionadas con las áreas temáticas del *Corpus Tècnic* del IULA: Genoma, Farmacogenómica, Neurociencia, Enfermedades, Eugenesia, Biotecnología, Diferenciación, Inmunología, Investigación genética, Estructura interna,

Ingeniería genética, Filogenia. Sin embargo, los datos muestran las tendencias de los sintagmas al principio de polaridad de Quirk *et al* (1985). Es decir, la clase *attribute* se encuentra en la mayoría de casos a la izquierda del sintagma (25 casos contra 16 en posición prenuclear) y se refiere a adjetivos predicativos o generales como se observa en los ejemplos de 15.

15. central, complex, effective, green, high, individual, large, multiple, natural, red, smooth, white.

De las 25 ocurrencias, sólo 3 no se podrían considerar como adjetivos generales (*adipose, contiguous, control*). Así, las características menos estables o más subjetivas tienden a ir más a la izquierda del sintagma como lo plantea Quirk *et al* (1985: 1341).

Obsérvese igualmente en la tabla 4 que la clase *adj.all* aumenta de frecuencia a medida que la posición se acerca más al núcleo. En la primera posición ocupa el puesto 13 con un 1,72% de todas las ocurrencias, en la segunda posición ocupa el puesto 8 con un 3,45% y en la tercera posición es la clase más frecuente con un 30,77%.

Sin embargo, existen otros adjetivos no etiquetados como *adj.all* en cada posición que son *adj.pert*, lo que muestra que el promedio de adjetivos aumenta a medida que la premodificación se aleja del núcleo a la izquierda, como se explica a continuación.

Primera posición	Porcentaje	Segunda posición	Porcentaje	Tercera posición	Porcentaje
noun.substance	22,41	noun.substance	21,12	adj.all	30,77
noun.body	19,83	noun.animal	15,95	noun.substance	15,38
noun.attribute	7,76	noun.body	12,5	noun.person	7,69
noun.animal	7,33	noun.attribute	10,34	noun.object	7,69
noun.cell	6,03	noun.state	4,31	noun.cognition	7,69
noun.state	4,74	verb.change	4,31	noun.attribute	7,69
noun.group	4,31	adj.all	3,45	noun.artifact	7,69
noun.artifact	3,02	noun.artifact	3,02	noun.animal	7,69
noun.event	3,02	noun.object	3,02	noun.act	7,69
noun.process	3,02	noun.process	2,16		
noun.act	2,59	noun.act	1,72		
adj.all	1,72	noun.cell	1,72		
noun.location	1,72	verb.cognition	1,72		
noun.phenomenon	1,72	noun.cognition	1,29		
noun.cognition	1,29	noun.group	1,29		
noun.quantity	0,86	noun.location	0,86		
verb.communication	0,86	noun.person	0,86		
noun.object	0,43	adv.all	0,43		
noun.shape	0,43	noun.food	0,43		
verb.cognition	0,43	noun.phenomenon	0,43		
verb.contact	0,43	noun.plant	0,43		
notWN	6,03	noun.relation	0,43		
		noun.shape	0,43		
		verb.communication	0,43		
		verb.contact	0,43		
		verb.motion	0,43		
		verb.perception	0,43		
		notWN	6,03		

Tabla 4: Clases semánticas por posiciones en la premodificación.

Así, hay 57 palabras (de 232) en la primera posición con adjetivos con un promedio de 4,07 (una de cada cuatro palabras es adjetivo), de los cuales 33 son *adj.pert* y 24 *adj.all*, como se ven en los ejemplos de 16.

16. artificial, auditory, binding, bovine, cellular, chromosomal, colorectal, congenital, cytogenetic, deleterious, dense, dependent, dominant, epistatic, epithelial, fetal, fluorescent, genetic, genomic, glial, heavy, human, lysosomal, mitochondrial, modern, molecular, monoclonal, negative, nervous, neuronal, nonradioactive, nuclear, ovine, partial, paternal, peripheral, placental, polymorphic, red, sexual, simplex, specific, transgenic, ventricular.

En la segunda posición, hay 120 palabras (de 232) con adjetivos, participios o adverbios deadjetivales con un promedio de 1,93, de las cuales 37 son *adj.pert* y 83 *adj.all* como se ven en los ejemplos de 17. Si bien se ha dicho en la metodología que los participios se separarían para los análisis sintácticos, es necesario incluirlos en los adjetivos para analizar aspectos semánticos y así obtener más regularidades.

17. aberrant, adipose, advanced, amino, anatomically, antiviral, apoptotic, automated, autosomal, bovine, central, chemical, circulating, columnar, complex, contiguous, critical, cytoplasmic, double, effective, embryonic, endothelial, environmentally, epidermal, epithelial, erythroid, eukaryotic, expected, false, fetal, genetic, genomic, green, growing, high, horizontal, human, immature, inbred, increased, individual, inner, known, large, masked, maternal, medial, meiotic, metastatic, mitochondrial, molecular, multiple, natural, neuronal, normal, outer, paramagnetic, parental, prandial, published, pulverized, ragged, red, reduced, repeated, repressed, smooth, stellate, total, verified, white.

En la tercera posición, hay 7 palabras (de 13 en total) con adjetivos con un promedio de 1,85, de las cuales 5 son *adj.pert* y 2 como *adj.all* se ve en los ejemplos de 18.

18. somatic, fetal, somatic, unequal, somatic, white, American.

Así, el promedio de adjetivos en cada posición tiende a aumentar de derecha a izquierda, es decir, que a medida que el premodificador se aleja del núcleo existen más probabilidades de ser adjetivo como lo reflejan los datos.

6.4.2.2 Clases semánticas de acuerdo con la categoría léxica

Si se observan las clases semánticas de la premodificación de acuerdo con la categoría léxica (N, Adj y PP) se evidencian diferencias en las clases semánticas que predominan en cada categoría.

Sustantivos	Porcentaje	Adjetivos	Porcentaje	Participios	Porcentaje
noun.substance	32,37	noun.animal[64]	20,63	verb.change	40
noun.body	16,91	noun.attribute	18,13	verb.cognition	16
noun.animal	7,91	noun.body	16,88	verb.communication	12
noun.cell	5,76	adj.all	10	noun.act	8
noun.attribute	4,32	noun.group	6,88	noun.process	8
noun.state	4,32	noun.state	5,63	verb.contact	8
noun.artifact	3,6	noun.artifact	3,13	verb.motion	4
noun.object	3,24	noun.substance	3,13	verb.perception	4
noun.event	2,52	noun.phenomenon	2,5		
noun.act	2,16	noun.process	2,5		
noun.cognition	2,16	noun.act	1,88		
noun.process	2,16	noun.person	1,88		
noun.location	1,08	noun.cell	1,25		
noun.group	0,72	noun.shape	1,25		
noun.quantity	0,72	noun.cognition	0,63		
noun.phenomenon	0,36	noun.food	0,63		
noun.plant	0,36	noun.location	0,63		
notWN	9,35	noun.relation	0,63		
		verb.cognition	0,63		
		notWN	1,25		

Tabla 5: Clases semánticas por categoría léxica en la premodificación.

En los sustantivos predomina la clase *substance* con un 32,37% mientras que en los adjetivos predomina la clase *animal* con un 20,63% y en los participios la clase *change* con un 40%. Luego, sigue la clase *body* (16,91%) en los sustantivos, la clase *attribute* (18,13%) en los adjetivos y la clase *verb.cognition* (16%) en los participios. La clase *body* también es igualmente frecuente en los adjetivos (16,88%).

Así, hay 134 registros de sustantivos con *substance* (32,37%) como se ve en los ejemplos de 19.

19. acid, agarose, albumin, amino, amyloid, androgen, APOE, buffer, calcium, carbon, cDNA, CTD, cytokine, DNA, dodecyl, endonuclease, enzyme, FHIT, fiber, glucose, guanylate, H1, histone, HLA-DR2, HMGIC, HUVEC, I, Igf2r, kinase, mannose, molecule, monolayer, mRNA, MRP, MUC7, myosin, NGF, nucleoside, nucleotide, oligosaccharide, phosphatase, phosphate, platelet,

[64] Aparece la etiqueta de *noun* porque son adjetivos relacionales (*pertainyms*)

platelet, pol, polyacrylamide, polymerase, polypeptide, potassium, product, promoter, protein, residues, RNA, sodium, sulfate, sulphatase, Taq, tau, tetrachloride, transcript, tyrosine.

Sin embargo, sólo hay 5 ocurrencias (3,13%) como adjetivos con la clase *substance* como se observa en 20. En cambio, hay 112 (10%) de adjetivos generales (*adj.all*) como en 21.

20. molecular, chemical.

21. aberrant, artificial, central, complex, critical, dense, dependent, dominant, double, effective, false, heavy, high, horizontal, human, immature, individual, inner, large, modern, multiple, natural, negative, nervous, normal, outer, partial, paternal, peripheral, red, smooth, specific, total, unequal, white.

A partir de estas observaciones, puede afirmarse que existen tendencias en el uso de determinadas clases semánticas de acuerdo con la categoría léxica, cuestión que puede afectar los aspectos denominativos en el caso de términos. Es decir, que para denominar sustancias, el sustantivo es la categoría léxica por excelencia al menos para esta área del conocimiento. Es difícil generalizar este asunto para otras áreas puesto que los aspectos semánticos están íntimamente ligados al área en cuestión. Esto puede observarse en otros estudios que han tenido las mismas restricciones semánticas (Oster 2005, 221).

6.4.3 Patrones semánticos obtenidos de WordNet 2.1

En la tabla 6, se presentan los patrones semánticos de más de 2 de frecuencia (28 patrones) que se han obtenido de WordNet 2.1. Para una muestra de 232 sintagmas seleccionados, existen 182 patrones diferentes (una media de 1,27), es decir, casi un patrón por cada sintagma. De estos 182 patrones, hay 154 patrones de una sola ocurrencia. Por tanto, puede afirmarse que existen muy

pocas regularidades en los patrones semánticos salvo por los patrones más frecuentes.

A continuación, se presentan los patrones semánticos más frecuentes conjuntamente con los patrones sintácticos con los que se correlacionan.

El patrón *animal notWN body* es el más frecuente con 6 ocurrencias y se correlaciona sintácticamente con el patrón Adj N N en 5 ocurrencias, como se observa en los ejemplos de 22.

22. human X chromosome, human IL11RA gene, human PAR genes, human IGF2R gene, human TnTf gene.

El patrón *animal substance body* es también el más frecuente con 6 ocurrencias y se correlaciona sintácticamente con el patrón Adj N N en 5 ocurrencias, como se presenta en los ejemplos de 23.

23. human NGF gene, human MRP genes, human tau gene, human HMGIC gene, human APOE gene.

Luego sigue el patrón *substance substance process* con 5 ocurrencias, y se correlaciona sintácticamente con el patrón N N N en 4 ocurrencias, como se ve en los ejemplos de 24.

24. H1 kinase activity, amino acid replacements, tyrosine kinase activity, CTD phosphatase activity.

El patrón *body body substance* tiene 4 ocurrencias, 2 de ellas tienen como patrón sintáctico N Adj N, como se indica en los ejemplos de 25.

25. lung lysosomal enzymes, kidney lysosomal enzymes.

El patrón *substance substance substance* tiene 4 ocurrencias y todas se correlacionan sintácticamente con el patrón N N N, como en los ejemplos de 26.

26. amino acid residues, sodium dodecyl sulfate, histone H1 kinase, Taq DNA polymerase.

Patrón semántico	Frecuencia	%
animal notWN body	6	2,59
animal substance body	6	2,59
substance substance process	5	2,16
body body substance	4	1,72
substance substance substance	4	1,72
adj.all attribute act	3	1,29
adj.all body body	3	1,29
animal animal body	3	1,29
animal body substance	3	1,29
attribute body cell	3	1,29
change body act	3	1,29
substance substance attribute	3	1,29
animal body group	2	0,86
animal process body	2	0,86
attribute state body	2	0,86
body attribute act	2	0,86
body body body	2	0,86
body location substance	2	0,86
body process body	2	0,86
group substance object	2	0,86
notWN attribute act	2	0,86
notWN body cell	2	0,86
object artifact quantity	2	0,86
substance act animal	2	0,86
substance artifact process	2	0,86
substance attribute object	2	0,86
substance substance cognition	2	0,86
substance substance group	2	0,86

Tabla 6: Patrones semánticos obtenidos con Wordnet 2.1.

Dentro de los patrones que tienen tres ocurrencias, cabe destacar el patrón *attribute body cell* que tiene el mismo patrón sintáctico, Adj N N, como se muestra en los ejemplos de 27.

27. white blood cells, smooth muscle cells, red blood cells.

Como antes se explicó, las clases semánticas que predominan y por extensión, los patrones semánticos son un claro reflejo del área temática del corpus. Así que los patrones semánticos, conjuntamente con los patrones sintácticos con los que se correlacionan, pueden ser un factor decisivo para la detección de unidades candidatas a término en un área del conocimiento determinada. Sin embargo, los patrones semánticos no pueden extrapolarse a otras áreas del conocimiento mientras que esto es posible con los patrones sintácticos como se ha visto en las tendencias presentadas en el corpus lexicográfico.

6.4.4 Patrones semánticos en la premodificación según WordNet 2.1

A pesar de las pocas regularidades que se encuentra en los patrones completos, se observan más regularidades al nivel de la premodificación como se observa en la tabla 7.

De los 116 patrones semánticos totales, 72 presentan una sólo ocurrencia, 14 con dos ocurrencias, 17 con tres ocurrencias y 4 con cuatro ocurrencias. Además, se encuentran patrones desde 5 ocurrencias hasta 20 ocurrencias. Los patrones semánticos de la premodificación con más de 5 ocurrencias son *substance substance* con el 8,62% (20 ocurrencias), *animal substance* con el 3,45% (8), *body body* con el 3,45% (8), *animal notWN* con el 3,02% (7), *attribute body* con el 2,59% (6), *substance body* con el 2,59% (6), *animal body* con el 2,16% (5) y *change body* con el 2,16% (5). En total representan el 28,04% de toda la premodificación. Igualmente, todos los patrones semánticos presentados en la tabla 7 representan el 56,85% de toda la premodificación.

Patrón semántico de la premodificación	Frecuencia	Porcentaje
substance substance	20	8,62
animal substance	8	3,45
body body	8	3,45
animal notWN	7	3,02
attribute body	6	2,59
substance body	6	2,59
animal body	5	2,16
change body	5	2,16
animal animal	4	1,72
attribute state	4	1,72
attribute substance	4	1,72
substance attribute	4	1,72
act animal	3	1,29
adj.all attribute	3	1,29
adj.all body	3	1,29
adj.all cell animal	3	1,29
animal group	3	1,29
animal process	3	1,29
body attribute	3	1,29
body cell	3	1,29
body location	3	1,29
body substance	3	1,29
cognition substance	3	1,29
group substance	3	1,29
notWN cell	3	1,29
object substance	3	1,29
state notWN	3	1,29
substance artifact	3	1,29
substance event	3	1,29

Tabla 7: Patrones semánticos en la premodificación según WordNet 2.1.

En cuanto al patrón semántico *substance substance* todos los casos pertenecen al patrón N N N aunque con diferentes núcleos. Con el patrón *animal substance*, 6 casos pertenecen al patrón Adj N N y 2 casos a N N N. En el patrón *body body* hay 3 casos que pertenecen al patrón Adj N N, 2 casos a N Adj N, 2 casos a N N N y 1 caso a Adj Adj N. El patrón *animal notWN* tiene 5 casos de los 7 con el patrón Adj N N y 2 casos con N N N. El patrón *attribute body* está representado por dos patrones Adj Adj N y Adj N N con 2 y 4 ocurrencias, respectivamente. El patrón *substance body* tiene dos patrones N N N con 5 ocurrencias y N JA N con una. El patrón *animal body* tiene también dos patrones Adj N N con 3 ocurrencias y Adj Adj N con 2. Finalmente, el patrón

change body tiene los patrones PP N N y PP Adj N con 3 y 2 ocurrencias, respectivamente.

De los anteriores datos, puede deducirse que los patrones más comunes en la premodificación son parte de los patrones sintácticos más frecuentes N N N, Adj N N, seguidos de Adj Adj N y N Adj N y en menor medida, PP N N, PP Adj N. Todos ellos están entre los 7 más frecuentes del corpus de análisis en inglés y entre los 8 más frecuentes del corpus lexicográfico del inglés. Así, puede afirmarse que estas tendencias son generales a este tipo de discurso especializado pero no es posible afirmar esto para otros tipos de discurso en el aspecto semántico ya que este está muy ligado al área temática y lo más probable es que estos patrones varíen de área en área temática como ya se explicó. Aún así, puede aseverarse que existen estructuras semánticas que subyacen a ciertos tipos de sintagmas en cada área temática como se ha visto en este apartado.

Obsérvese que los patrones Adj N N, N N N y N Adj N son los que están presentes en las estructuras semánticas más frecuentes y a su vez son los tres patrones más frecuentes tanto en el corpus de análisis como en el corpus lexicográfico (1, 2 y 4, respectivamente en ambos corpus). En este sentido, puede afirmarse que las estructuras más frecuentes tienden a estar correlacionadas semántica y sintácticamente. Una de las estructuras sintácticas más frecuentes Adj Adj N en ambos corpus (3ra) no presenta tanta correlación con las semánticas. Sólo presenta dos regularidades *animal animal body* y *animal body substance*, ambas con 2 ocurrencias. Las otras 23 ocurrencias de este patrón tienen igual cantidad de patrones semánticos. Como es de esperarse, muchas de las clases semánticas de la premodificación son *adj.all*, *attribute*, *shape* y *state*, lo que también muestra el tipo de clase que subyace a un patrón como Adj Adj N.

A pesar de la alta frecuencia presentada por los patrones Adj N N, N N N y N Adj N entre los patrones semánticos, es importante mostrar que estos

patrones sintácticos también presentan otros patrones semánticos diferentes de los expuestos.

El patrón N N N presenta también del patrón *substance substance proces*, el patrón *substance substance substance* con las mismas 4 ocurrencias y el patrón *substance substance attribute* con 3 ocurrencias. En estos tres patrones predomina una premodificación con la clase *substance substance* como sucede igualmente con otros dos patrones más de 2 ocurrencias (*substance substance cognition* y *substance substance group*). Además, este patrón presenta 6 patrones con 2 ocurrencias y 53 con una 1 sola ocurrencia.

El patrón Adj N N presenta además del patrón *animal notWN body* con 5 ocurrencias, los patrones *animal substance body* con las mismas ocurrencias y *adj.all attribute act* y *attribute body cell*, ambos con 3 ocurrencias. Además, este patrón sintáctico presenta 6 patrones semánticos de 2 ocurrencias y 38 de una sola ocurrencia.

Finalmente, el patrón N Adj N presenta además del patrón semántico *body body substance* con 2 ocurrencias, el patrón *substance attribute object* con 2 ocurrencias. El resto de patrones son de una ocurrencia.

Así, puede decirse de este apartado que existen inconsistencias en el etiquetaje de WordNet en cuanto que hay palabras que pueden pertenecer a dos clases semánticas como sucede con *enzyme* que pueden estar asignadas a *body* y a *substance* y que podrían estar etiquetadas y jerarquizadas desde ambas sin que existan ningún problema conceptual o cognitivo ya que, como lo plantea Cabré (2002), la poliedricidad de los términos permite analizarlos desde diferentes perspectivas y es en el marco de un área temática que adquieren su valor especializado:

> "... los términos, que son las unidades del campo de conocimiento llamado terminología, se pueden analizar desde perspectivas diferentes y, en tanto que objetos poliédricos, pueden participar de su campo de estudio y convertirse en

parte central del objeto de análisis y de su teorización. Desde la lingüística, se puede elaborar perfectamente una teoría de los términos en la que éstos se describen como unidades de forma y contenido que, utilizados en determinadas condiciones discursivas, adquieren un valor especializado."

Igualmente, las clases y patrones encontrados permiten observar la relación entre los sintagmas y su área temática. Por tanto, es de esperar que se obtengan estos resultados debido al área temática pero si se combinan con la descripción formal pueden ser útiles en los procesos de extracción de términos, en especial.

6.4.5 Análisis de las clases semánticas de los núcleos según UMLS

Como se dijo en 3.2.1.4, UMLS es un conjunto de recursos léxicos que facilita el desarrollo de los sistemas informáticos para que "entiendan" el lenguaje de la biomedicina y la salud y, por eso, presentan un nivel de granularidad muy superior a WordNet. Por tanto, se espera que los resultados sean más precisos y generalizables que los presentados desde el etiquetaje con WordNet.

En los núcleos se encontraron 58 tipos semánticos de los 135 totales. De estos, 43 son de una frecuencia menor a 5 ocurrencias y representan el 34,4%.

Clase semántica UMLS	Frecuencia	Porcentaje
Gene or Genome	22	9,48
Biologically Active Substance	21	9,05
Functional Concept	14	6,03
Cell	13	5,6
Quantitative Concept	11	4,74
Idea or Concept	10	4,31
Spatial Concept	9	3,88
Cell Component	8	3,45
Research Activity	7	3,02
Tissue	7	3,02
Genetic Function	6	2,59
Laboratory Procedure	6	2,59

Disease or Syndrome	5	2,16
Qualitative Concept	5	2,16
Mammal	4	1,72
Organism Function	4	1,72
Substance	4	1,72
Body Part, Organ, or Organ Component	3	1,29
Body Substance	3	1,29
Cell or Molecular Dysfunction	3	1,29
Enzyme	3	1,29
Intellectual Product	3	1,29
Medical Device	3	1,29
Molecular Function	3	1,29
Nucleotide Sequence	3	1,29
Phenomenon or Process	3	1,29
Mental Process	2	0,86
Natural Phenomenon or Process	2	0,86
Nucleic Acid, Nucleoside, or Nucleotide	2	0,86
Occupational Activity	2	0,86
Organic Chemical	2	0,86
Organism Attribute	2	0,86
Pharmacologic Substance	2	0,86
Research Device	2	0,86
Therapeutic or Preventive Procedure	2	0,86
Virus	2	0,86
Amino Acid Sequence	1	0,43
Amino Acid, Peptide, or Protein	1	0,43
Biomedical Occupation or Discipline	1	0,43
Biomedical or Dental Material	1	0,43
Carbohydrate	1	0,43
Chemical Viewed Structurally	1	0,43
Clinical Attribute	1	0,43
Clinical Drug	1	0,43
Congenital Abnormality	1	0,43
Finding	1	0,43
Food	1	0,43
Hazardous or Poisonous Substance	1	0,43
Hormone	1	0,43
Human	1	0,43
Inorganic Chemical	1	0,43
Manufactured Object	1	0,43
Neoplastic Process	1	0,43
Pathologic Function	1	0,43
Sign or Symptom	1	0,43
Social Behavior	1	0,43
not found	8	3,45
ok	1	0,43

Tabla 8: Clases semánticas del núcleo según UMLS.

Como puede verse en la tabla 8, la clase semántica más frecuente en el corpus de análisis es *Gene or Genome* con 9,48%, seguida de *Biologically Active Substance* con 9,05% y *Functional Concept* con 6,03%, como puede verse en los ejemplos 28, 29 y 30. Estas tres clases semánticas representan el 24,56% de todos los núcleos de la muestra semántica.

28. allele, gene, library.

29. albumin, buffer, collagen, DNA, mRNA, plasmids, protein, RNA.

30. collection, deficiency, domain, factor, fragment, reaction, replacements, results, shifts, system, transfer, turnover.

Los ejemplos de 28 y 29 muestran palabras relacionadas con sus respectivas clases y en el caso de 30, la clase *Functional Concept* se refiere a "*A concept which is of interest because it pertains to the carrying out of a process or activity*". Por tanto, es una clase muy amplia que no permite mayores restricciones o generalizaciones. Sin embargo, el ejemplo *fragment* puede pertenecer a *Gene or Genome* si se tiene en cuenta que UMLS es una ontología especializada en ciencias de la salud.

Posteriormente siguen las clases *Cell* con 5,6%, *Quantitative Concept* con 4,74% e *Idea or Concept* con 4,31%, como se observa en los ejemplos de 31, 32 y 33.

31. cell, platelet.

32. count, difference, distance, kit, mass, number, weight.

33. activity, chain, cluster, death, product, strain.

Hay un 3,45% de palabras que no se encontraron en UMLS y por tanto se etiquetaron como *not found*. Dichas unidades son palabras generales que puede

aparecer en la clase *Functional Concept* y otras son del ámbito del genoma y pueden etiquetarse como *Cell Component*, como se puede ver en los ejemplos de 34.

34. association, changes, locus, pocket, transition.

Si se comparan los resultados de WordNet contra los de UMLS, se refrenda lo expuesto por Burgun/Bodenreider (2001: 77):

> "Only 2% of the domain-specific concepts from UMLS were found in WordNet, but 83% of the domain-specific concepts from WordNet were found in the UMLS."

Así es posible obtener clases más específicas en UMLS como *enzyme*, *hormone, amino acid* que en WordNet se pueden agrupar en *substance*. Sin embargo, en UMLS esta clase semántica también está en un solo grupo semántico CHEM (*Chemicals&Drugs*). En este sentido, con UMLS puede lograrse dos aspectos importantes: obtener más generalizaciones con los grupos semánticos y lograr más precisión con los tipos semánticos. Con WordNet, solo se logra más generalización pero se pierde precisión semántica. Para efectos de extracción terminológica y etiquetaje en bancos de datos, UMLS sólo sería útil en el ámbito de la medicina mientras que WordNet, al ser una ontología general puede ser útil en muchos campos del conocimiento como medicina, economía, geología, arquitectura, transporte, etc. WordNet contiene 42 áreas de conocimiento aunque desarrolladas de manera diferente. Por ejemplo, el área de psicología tiene 3.405 *synsets* mientras que veterinaria tiene 92 (Magnini *et al* 2002: 363).

6.4.6 Análisis de las clases semánticas de la premodificación según UMLS

En la premodificación se encontraron 62 tipos semánticos de los 135 totales. De estos, 39 son de una frecuencia menor a 5 ocurrencias y representan el 17,22%.

Como puede verse en la tabla 8, la clase semántica más frecuente en el corpus de análisis es *Amino Acid, Peptide, or Protein* con 11,39%, seguido de *Qualitative Concept* con 7,59% y *Functional Concept* con 6,75% como puede verse en los ejemplos de 35, 36 37. Estas tres categorías semánticas representan el 32,06% de todos los núcleos.

35. A, albumin, amino, amyloid, apoE, collagen, cytokine, endonuclease, FHIT, H1, histone, HMGIC, HUVEC, I, Igf2r, insulin, kinase, MRP, MUC7, myosin, NGF, phosphatase, polymerase, polypeptide, protein, receptor, restriction, sulphatase, Taq, tyrosine.

36. aberrant, advanced, apoptotic, artificial, complex, congenital, critical, dense, effective, erythroid, false, green, heavy, high, known, marker, molecular, normal, paternal, red, reduced, smooth, specific, striated, Type, unequal, white.

37. anatomically, auditory, automated, cellular, circulating, domain, dominant, double, endogenous, epithelial, fetal, fragment, genetic, mitochondrial, natural, negative, nervous, repeated, sequencing, shift.

Posteriormente siguen las clases *Nucleic Acid, Nucleoside, or Nucleotide* con 5,91%, *Cell* con 5,06% y *Human* con 4,85% como se ve en los ejemplos 37, 39 y 40.

38. cDNA, DNA, genomic, guanylate, nucleoside, nucleotide, plasmid, promoter, start, tau, transcript.

39. adipocyte, cell, leukocytic, platelet, polarized, somatic, T.

40. human, individual.

Clase semántica UMLS	Frecuencia	Porcentaje
Amino Acid, Peptide, or Protein	54	11,39
Qualitative Concept	36	7,59
Functional Concept	32	6,75
Nucleic Acid, Nucleoside, or Nucleotide	28	5,91
Cell	24	5,06
Human	23	4,85
Spatial Concept	19	4,01
Gene or Genome	14	2,95
Quantitative Concept	14	2,95
Body Part, Organ, or Organ Component	12	2,53
Idea or Concept	11	2,32
Mammal	11	2,32
substance	8	1,69
Carbohydrate	6	1,27
Disease or Syndrome	6	1,27
Laboratory Procedure	6	1,27
Tissue	6	1,27
Body Location or Region	5	1,05
Body Substance	5	1,05
Cell Component	5	1,05
Organism	5	1,05
Amino Acid Sequence	4	0,84
Biologically Active	4	0,84
Cell or Molecular Dysfunction	4	0,84
Genetic Function	4	0,84
Inorganic Chemical	4	0,84
Animal	3	0,63
Antibiotic	3	0,63
Biomedical or Dental	3	0,63
Clinical Attribute	3	0,63
Immunologic Factor	3	0,63
Organism Function	3	0,63
Population Group	3	0,63
Virus	3	0,63
Cell Function	2	0,42
Chemical	2	0,42
Element, Ion, or Isotope	2	0,42
Experimental Model of Disease	2	0,42
Finding	2	0,42
Machine Activity	2	0,42
Natural Phenomenon or Process	2	0,42
Organism Attribute	2	0,42
Pharmacologic Substance	2	0,42
Research Device	2	0,42
Social Behavior	2	0,42
Temporal Concept	2	0,42
Bacterium	1	0,21
Biomedical Occupation or Discipline	1	0,21
Embryonic Structure	1	0,21

Family Group	1	0,21
Fungus	1	0,21
Indicator, Reagent, or Diagnostic Aid	1	0,21
Intellectual Product	1	0,21
Lipid	1	0,21
Medical Device	1	0,21
Neoplastic Process	1	0,21
Occupational Activity	1	0,21
Phenomenon or Process	1	0,21
Research Activity	1	0,21
Therapeutic or Preventive Procedure	1	0,21
not found	62	13,08

Tabla 9: Clases semánticas de la premodificación en UMLS.

Hay un 13,08% de palabras de la premodificación que no se encontraron en UMLS y, por tanto, se etiquetaron como *not found*. Dichas unidades son, en general, términos o parte de ellos, que puede aparecer en la clase *Functional Concept* y otras son del ámbito del genoma y pueden etiquetarse como *Cell Component* como se puede ver en el ejemplo 41.

41. ArG, BD, binding, biotinylated, CaCo2, calf, chromosomal, corpuscle, deleterious, ELT-3, endothelial, environmentally, etl2, eukaryotic, extinguisher, genomic, growing, helper, highly, HPS, immunoreactive, K2, Kozak, modern, MUL, multipoint, nonradioactive, PAR, paramagnetic, parental, polymorphic, prandial, pseudocontact, published, pulverized, ragged, repressed, SDS, sexual, simplex, subunit, TA, TATAA, TnTf, variance.

6.4.7 Patrones semánticos obtenidos de UMLS

Debido a la cantidad de tipos semánticos que tiene UMLS y, a que tal cantidad reduce la posibilidad de hacer generalizaciones sobre los patrones semánticos, se han empleado los grupos semánticos y no los tipos semánticos expuestos en §2. UMLS tiene 15 grupos semánticos que permiten agrupar los tipos semánticos. Para poder obtener regularidades desde el punto de vista de los patrones se han mapeado los 136 tipos semánticos a sus respectivos grupos semánticos.

En la tabla 10, se presentan los patrones semánticos que se han obtenido de los grupos semánticos de UMLS. Se presentan todos los patrones que tienen una frecuencia igual o mayor que 2 (40 patrones).

Para una muestra de 232 sintagmas seleccionados existen 149 patrones diferentes (una media de 1,55), es decir, casi un patrón por cada sintagma y medio. De estos 149 patrones, hay 110 patrones de una sola ocurrencia. Por tanto, puede verse que al igual que con los patrones de WordNet 2.1, existen muy pocas regularidades en los patrones semánticos aún si se mapean los tipos semánticos a los grupos semánticos que en teoría tienen.

A continuación, se presentan los patrones semánticos más frecuentes con los patrones sintácticos que ellos representan.

De los 149 patrones semánticos totales, 110 presentan una sola ocurrencia, 17 con dos ocurrencias, 14 con tres ocurrencias y 3 con cuatro ocurrencias. Además, se encuentran patrones desde 5 ocurrencias hasta 8 ocurrencias.

El patrón CHEM CHEM CONC (*Chemicals&Drugs Chemicals&Drugs Concepts&Ideas*) es el más frecuente con 8 ocurrencias (3,45%) y se correlaciona sintácticamente con el patrón N N N en 8 ocurrencias, como se aprecia en los ejemplos de 42.

42. H1 kinase activity, amino acid replacements, tyrosine kinase activity, amino acid level, calcium phosphate method, amino acid differences, CTD phosphatase activity, CTD phosphate turnover.

Patrón UMLS	Frecuencia	Porcentaje
CHEM CHEM CONC	8	3,45
LIVB CHEM GENE	6	2,59
CHEM CHEM CHEM	5	2,16
CHEM CONC CONC	5	2,16
CONC ANAT ANAT	5	2,16
CONC CONC CONC	5	2,16
CHEM CHEM PHYS	4	1,72
CONC CHEM CHEM	4	1,72
LIVB CONC ANAT	4	1,72
ANAT ANAT ANAT	3	1,29
ANAT ANAT CHEM	3	1,29
ANAT ANAT CONC	3	1,29
CHEM CHEM GENE	3	1,29
CHEM CHEM PROC	3	1,29
CONC ANAT DEVI	3	1,29
CONC CONC ANAT	3	1,29
CONC CONC DISO	3	1,29
CONC LIVB ANAT	3	1,29
LIVB GENE GENE	3	1,29
LIVB NotF GENE	3	1,29
NotF CHEM CONC	3	1,29
NotF NotF CONC	3	1,29
ok ok ACTI	3	1,29
ANAT CONC ANAT	2	0,86
CHEM ACTI CONC	2	0,86
CHEM CHEM DEVI	2	0,86
CHEM CHEM NotF	2	0,86
CHEM DISO LIVB	2	0,86
CHEM GENE PHYS	2	0,86
CHEM NotF CHEM	2	0,86
CHEM NotF NotF	2	0,86
CONC ANAT CONC	2	0,86
CONC ANAT DISO	2	0,86
CONC GENE CONC	2	0,86
GENE GENE CHEM	2	0,86
LIVB ANAT CHEM	2	0,86
NotF ANAT ANAT	2	0,86
NotF ANAT OBJC	2	0,86
NotF CHEM PHYS	2	0,86
NotF PHYS CONC	2	0,86

Tabla 10: Patrones semánticos en UMLS.

El patrón LIVB CHEM GENE (*Living Beings Chemicals&Drugs Genes&MolecularSequences*) tiene 6 ocurrencias (2,59%) y se correlaciona sintácticamente con el patrón Adj N N en 5 ocurrencias, como se observa en los ejemplos de 43.

43. human NGF gene, human MRP genes, human tau gene, human HMGIC gene, human APOE gene.

El patrón CHEM CHEM CHEM *(Chemicals&Drugs Chemicals&Drug Chemicals&Drug)* tiene 5 ocurrencias (2,16%) y se correlaciona sintácticamente con el patrón N N N en 5 ocurrencias, como se ve en los ejemplos de 44.

44. amino acid residues, plasmid DNA vaccines, sodium dodecyl sulfate, histone H1 kinase, Taq DNA polymerase.

El patrón CHEM CONC CONC (*Chemicals&Drugs Concepts&Ideas Concepts&Ideas*) tiene 5 ocurrencias (2,16%) y se correlaciona sintácticamente con el patrón N Adj N en 2 ocurrencias, como se presenta en los ejemplos de 45.

45. glucose specific activity, myosin heavy chain.

El patrón CONC ANAT ANAT (*Concepts&Ideas Anatomy Anatomy*) tiene 5 ocurrencias (2,16%) y se correlaciona sintácticamente con el patrón Adj N N en 3 ocurrencias, como se muestra en los ejemplos de 46.

46. white blood cells, smooth muscle cells, red blood cells.

El patrón CONC CONC CONC (*Concepts&Ideas Concepts&Ideas Concepts&Ideas*) tiene 5 ocurrencias (2,16%) y se correlaciona sintácticamente con el patrón Adj|PP Adj N en 5 ocurrencias, como aprecia en los ejemplos de 47.

47. central nervous system, large molecular weight, false negative results, natural genetic variation, expected molecular mass.

Estos patrones semánticos representan el 14,68% del total de ocurrencias.

6.4.8 Patrones semánticos en la premodificación según UMLS

A pesar de las pocas regularidades que se encuentra en los patrones completos, se observan muchas más regularidades al nivel de la premodificación que a nivel del patrón completo como se observa en la tabla 11.

Premodificación	Frecuencia	Porcentaje
CHEM CHEM	29	12,5
CONC ANAT	14	6,03
CONC CONC	14	6,03
ANAT ANAT	11	4,74
CONC CHEM	11	4,74
NotF CHEM	9	3,88
CHEM CONC	8	3,45
LIVB CHEM	7	3,02
CONC LIVB	6	2,59
CONC NotF	6	2,59
NotF ANAT	6	2,59
CHEM NotF	5	2,16
CONC DISO	5	2,16
LIVB CONC	5	2,16
LIVB NotF	5	2,16
ANAT CONC	4	1,72
LIVB ANAT	4	1,72
LIVB GENE	4	1,72
ok ok	4	1,72
ANAT ANAT LIVB	3	1,29
CHEM GENE	3	1,29
CHEM PHYS	3	1,29
CONC GENE	3	1,29
CONC PHYS	3	1,29
CONC PROC	3	1,29
LIVB DISO	3	1,29
LIVB LIVB	3	1,29
NotF CONC	3	1,29
NotF NotF	3	1,29

Tabla 11: Patrones semánticos en la premodificación según UMLS.

De los 66 patrones semánticos totales, 29 presentan una sólo ocurrencia, 8 con dos ocurrencias, 10 con tres ocurrencias y 4 con cuatro ocurrencias.

Además, se encuentran patrones desde 5 ocurrencias hasta 29 ocurrencias. Los patrones semánticos de la premodificación con más de 5 ocurrencias: CHEM CHEM con el 12,5% (29 ocurrencias), CONC ANAT con el 6,03% (14), CONC CONC con el 6,03% (14), ANAT ANAT con el 4,74% (11), CONC CHEM con el 4,74% (11), NotF CHEM con el 3,88% (9), CHEM CONC con el 3,45% (8), LIVB CHEM con el 3,02% (7),CONC LIVB con el 2,59% (6), CONC NotF con el (6) 2,59%, NotF ANAT con el 2,59% (6), CHEM NotF con el 2,16% (5), CONC DISO con el 2,16% (5), LIVB CONC con el 2,16% (5), LIVB NotF con el 2,16% (5). En total representan el 60,8% de toda la premodificación.

El patrón CHEM CHEM tiene 27 casos que pertenecen al patrón N N N aunque con diferentes núcleos. Con el patrón CONC ANAT, 12 casos pertenecen al patrón Adj N N y 2 casos a PP N N. En el patrón CONC CONC hay 10 casos que pertenecen al patrón Adj Adj N y 2 casos a PP Adj N. El patrón ANAT ANAT tiene 5 casos con el patrón Adj N N, 4 casos con N N N y 2 casos con N Adj N. El patrón CONC CHEM está representado por los patrones Adj N N con 6 ocurrencias y los patrones PP N N y N N N con 2 ocurrencias, respectivamente. Con el patrón NotF CHEM se pueden considerar dos patrones Adj N N con 6 ocurrencias y PP N N con 2 ocurrencias. El patrón CHEM CONC tiene el patrón N N N con 3 ocurrencias y los patrones N Adj N y Adj N N con 2 ocurrencias cada uno. El patrón LIVB CHEM tiene los patrones Adj N N con 5 ocurrencias y N N N con 2. El patrón CONC LIVB tiene los patrones Adj Adj N con 4 ocurrencias y PP|Adj N N con 2. El patrón CONC NotF tiene los patrones PP|Adj Adj N, Adj N N y Adv Adj N, todos con 2 ocurrencias. El patrón NotF ANAT tiene el patrón N N N con cuatro ocurrencias. El patrón CHEM NotF tiene los patrones N Adj N y N N N con dos ocurrencias cada uno. El patrón CONC DISO tiene el patrón Adj N N con cuatro ocurrencias. El patrón LIVB CONC tiene el patrón Adj Adj N con 3 ocurrencias. Finalmente, el patrón LIVB NotF tiene los patrones N N N y Adj N N, ambos 2 con ocurrencias.

De los anteriores datos, puede deducirse que los patrones más comunes en la premodificación son, de nuevo, parte de los patrones sintácticos más frecuentes N N N, Adj Adj N y Adj N N, seguidos de N Adj N, PP N N y PP Adj N

y, en menor medida, Adv Adj N. Al igual que las tendencias presentadas con los patrones de WordNet, todos ellos pertenecen a los 7 más frecuentes del corpus de análisis en inglés y entre los 8 más frecuentes del corpus lexicográfico del inglés.

6.5 Recapitulación

En este capítulo, se han presentado los resultados del análisis semántico del corpus de análisis en inglés en WordNet 2.1 y UMLS 2006.

1. Las clases semánticas más frecuentes en el núcleo en WordNet 2.1 son *noun.body* (18,53%), *noun.substance* (15,95%), *noun.act* (10,77%), *noun.group* (7,33%) y *noun.process* (7,33%). Estas cinco clases semánticas representan el 59,91% de todos los núcleos de la muestra. En UMLS, las clases semánticas más frecuentes son *Gene or Genome* (9,48%), *Biologically Active Substance* (9,05%), *Functional Concept* (6,03%), *Cell* (5,6%) y *Quantitative Concept* (4,74%). Estas cinco clases representan el 34,9% de todos los núcleos. Obsérvese que WordNet 2.1 tiene más capacidad de generalización pero UMLS presenta más granularidad ya que las clases de los núcleos están más repartidas entre las diferentes clases.

2. Las clases semánticas más frecuentes en la premodificación en WordNet son *noun.substance* (21,59%), *noun.body* (15,72%), *noun.animal* (11,53%), *noun.attribute* (9,01%) y, *noun.state* (4,4%). En UMLS las clases más frecuentes son *Gene or Genome* (9,48%), *Biologically Active Substance* (9,05%), *Functional Concept* (6,03%), *Cell* (5,6%) y *Quantitative Concept* (4,74%). También en la premodificación se presenta la misma tendencia de granularidad frente a WordNet en los núcleos ya que estas clases representan el 34,9% de toda la premodificación en UMLS contra el 62,25% en WordNet.

3. Los patrones más frecuentes en WordNet son *animal notWN body* (2,59%), *animal substance body* (2,59%), *substance substance process* (2,16%), *body body substance* (1,72%) y *substance substance substance* (1,72%). Estos patrones semánticos obtenidos a partir de WordNet representan el 10,78% de la muestra de análisis semántico. Los patrones más frecuentes en UMLS son

CHEM CHEM CONC (3,45%), LIVB CHEM GENE (2,59%), CHEM CHEM CHEM (2,16%), CHEM CONC CONC (2,16%) y CONC ANAT ANAT (2,16%). Estos patrones semánticos obtenidos de UMLS representan el 12,52% de toda la muestra de análisis en inglés. Puede verse que en ambos programas no es posible obtener muchas generalizaciones en cuanto a los patrones ya que cada patrón semántico no abarca a más del 3,5% de todas las ocurrencias en el mejor de los casos. Sin embargo, los patrones más frecuentes en ambos sistemas se correlacionan sintácticamente con los patrones superficiales más frecuentes tanto en el corpus de análisis en inglés como en el lexicográfico: N N N, Adj Adj N y Adj N N, N Adj N, PP N N y PP Adj N y en menor medida Adv Adj N. Puesto que los patrones semánticos tienen las clases semánticas más frecuentes, los patrones creados a partir de ellas y su asociación a los patrones superficiales más frecuentes muestra que son estas estructuras las más estables dentro de este estudio en todo sentido.

4. Los resultados reflejan lo "esperable" en cuanto a las clases semánticas puesto que el área temática de este estudio, el genoma, tiene involucradas estas clases. Por tanto, su aporte a este estudio es limitado. Sin embargo, el análisis realizado y la asociación que se ha hecho entre los patrones superficiales y los semánticos permiten saber que un uso adecuado entre los patrones y las clases semánticas de un área temática determinada, e.g., economía, y teniendo en cuenta que algunas de ellas están bien desarrolladas en determinadas ontologías, es posible trasladar los resultados de este estudio hacia campos de aplicación, como el etiquetaje de corpus, traducción automática, ontologías, extracción de terminología, lexicografía, etc.

5. Aun así es importante señalar las limitaciones de ambos sistemas y que se han esbozado antes. En primer lugar, existe una cantidad importante de palabras que no se encuentran en ambos sistemas: 16,53% en UMLS y 7,4% en WordNet 2.1. Esto puede afectar los resultados de cualquier estudio ya que sobrepasan el estándar del 5% de error y para subsanar esto, se debe entrenar el corpus y detectar el porcentaje potencial de unidades que no tiene el sistema, etiquetarlo manualmente y alimentar el sistema hasta reducirlo por debajo del

5%. Esto puede variar de área en área ya que en el caso de WordNet, la granularidad varía ostensiblemente como se ha comentado a final de §6.4.5.

En el caso de UMLS, al estar restringida a las ciencias de la salud es muy útil para trabajos como este, pero también tiene un porcentaje importante de palabras no encontradas. Como se ha visto aquí el uso de los grupos semánticos permite hacer más generalizaciones aunque disminuya la granularidad, en cuyo caso se deben evaluar las ventajas y desventajas para sacar el máximo provecho en el etiquetaje o emplear un etiquetaje doble en todos los casos.

Finalmente, debe tenerse en cuenta las inconsistencias en el etiquetaje de ambos sistemas que se han explicado para poder obtener más regularidades en los resultados.

7. Descripción y análisis semántico de los patrones en español

7.1 Introducción

En este capítulo se pretende caracterizar semánticamente los sintagmas nominales especializados extensos en español de modo que conjuntamente con los resultados sintácticos permitan interpretar estos sintagmas usando elementos lingüísticos obtenidos de las regularidades observadas.

Aunque no se pretende establecer las relaciones semánticas entre los diferentes constituyentes como lo ha hecho Oster (2005) se busca establecer patrones semánticos de los sintagmas empleando EuroWordNet 1.6. Estos patrones pueden permitir observar las tendencias semánticas en español, al menos para este campo del genoma.

Se ha elegido EuroWordNet para el español por varias razones. En primer lugar, es la versión europea de WordNet 1.6, lo que permite obtener resultados comparables en ambas lenguas en clases semánticas y los patrones obtenidos. En segundo lugar, es la ontología más empleada en todos los estudios relacionados con la lingüística y en la construcción de otros recursos lingüísticos como diccionarios, tesauros, etc. y el procesamiento del lenguaje natural. En tercer lugar, EuroWordNet es de acceso libre a través de Internet.

Aunque UMLS tiene algunos recursos para el español como Snomed, no es posible acceder a ellos vía Web para etiquetar las palabras y por eso, no se empleó en este capítulo.

7.2 Criterios y selección del corpus de análisis en español

Como se explicó en el capítulo de la metodología (§3.7), para el análisis semántico se seleccionó manualmente una muestra de los 8 patrones más frecuentes que representan el 78,36% a partir de la muestra del análisis sintáctico. Para ello, se seleccionó un 22[65]% de los sintagmas y se distribuyó proporcionalmente de acuerdo con su frecuencia como se hizo con la muestra sintáctica. Es decir, al patrón más frecuente le correspondían más sintagmas para el análisis semántico y al patrón menos frecuente se le asignaban menos sintagmas. Por ejemplo, el patrón N Prep N Adj es el más frecuente del corpus y le corresponden 31 sintagmas y los menos frecuentes son N Adj PP y N Prep Adj N y les corresponden 6 y 5 sintagmas de la muestra, respectivamente.

	Muestra sintáctica	Muestra semántica	Porcentaje
Español	1.055	200 SN	22%

Tabla 1: Muestra de sintagmas extraídos para el análisis semántico.

Al igual que en inglés, se seleccionaron los sintagmas para la muestra semántica siguiendo el mismo criterio que para el análisis de la muestra sintáctica, es decir, de mayor a menor frecuencia hasta completar la muestra que correspondía a cada patrón sintáctico.

[65] Aunque una muestra de 200 sintagmas puede parecer limitada, la decisión del 22% se basa en la selección de un 20% de la muestra de forma completa para cada patrón y por eso se ha llegado hasta ese porcentaje. Aunque está limitación se justifica por la dificultad de asignar manualmente hasta 6 etiquetas a cada uno de los tokens de estos sintagmas (un total de 606 tokens en cada sistema). Además, cada token se ha buscado, desambiguado en otros diccionarios, y observado en contexto para poder asignar la etiqueta de EuroWordNet de manera correcta. Para más información, véase §3.4 y §7.3 de este capítulo.

7.3 Metodología

La muestra seleccionada se etiquetó manualmente para un total de 606 tokens. Las categorías semánticas, presentadas en §3.2.2.3, se asignaron por separado en todos los núcleos con EuroWordNet. Posteriormente, se etiquetó la modificación de acuerdo con las categorías léxicas: sustantivos, adjetivos y adverbios, de modo que pudiera asegurarse la consistencia de etiquetaje dentro de cada categoría léxica en EuroWordNet. Si se encontraba más de una posibilidad de etiquetaje se asignaba el que correspondiera más al sintagma o, en caso de ambigüedad se dejaban las diferentes clases. En cualquier caso, la primera opción del etiquetaje se dejaba como la más próxima al significado del sintagma.

Los datos se tabularon en Statgraphics 5.1 en cuanto a las clases semánticas de los núcleos y la modificación en EuroWordNet. De allí, se seleccionaron las categorías más prototípicas de la muestra tanto para los núcleos como para la modificación.

Se obtuvo luego el conjunto de patrones semánticos más frecuentes y se compararon con los patrones sintácticos, de modo que pueda verse la relación entre los aspectos sintáctico y semántico y las tendencias que presentaban. Igualmente se analizó su frecuencia en cuanto a la posición dentro de la modificación y de acuerdo con la categoría léxica.

7.4 Resultados

7.4.1 Análisis de las clases semánticas de los núcleos según EuroWordNet

Como puede verse en la tabla 2, la clase semántica más frecuente en el corpus de análisis del español es *state* con 21%, seguida de *act* con 12% y *body* con 11%. Estas tres categorías semánticas representan el 44% de todos los núcleos.

Como puede verse en el ejemplo 1, los casos denotan enfermedades (anemia, leucemia), estados de enfermedades (anomalía, nivel, fiebre, grado) o estados en general (muerte, nivel).

1. agenesia, anemia, anomalía, artrosis, atrofia, candidiasis, carcinoma, cáncer, diabetes, dominio, enfermedad, existencia, fetopatía, fiebre, grado, insuficiencia, leucemia, muerte, nivel, poliposis, poliquistosis, tumor.

Igualmente, hay algunos casos que no se puede categorizar en esta clase. Por ejemplo, dominio debería pertenecer a la clase *cell* de acuerdo con el diccionario Espasa de medicina.

(Gen.) m. segmento, habitualmente pequeño, de DNA o de un polipéptido, que tiene una función o unas propiedades específicas.

En el ejemplo 2, se presentan casos de la clase *act*. Obsérvese que todos los sustantivos son deverbales.

2. análisis, cultivo, delito, diagnóstico, distribución, formación, hibridación, inclusión, reparación, respuesta, rotura, secuencia, tinción, transferencia, transmisión, índice.

En el caso de cultivo, debería estar bajo la clase *substance* o *cell* como se infiere de las definiciones de la DRAE y el Espasa si bien la categoría *act* es precisa para dicha palabra.

> (Biol. y Med.) población de microorganismos, células o tejidos así obtenidos.
>
> (Microbiol.) m. medio sólido o líquido en el que se ha propagado una población de un determinado tipo de microorganismo (o célula de un macroorganismo), como resultado de la previa inoculación de ese medio, seguida de una incubación.

En el caso de hibridación, se presenta una situación similar ya que pertenece más al ámbito de la biología como tal y podría clasificarse bajo *cell* como se observa de las definiciones del DRAE y Espasa de medicina.

> (Biol.) fusión de dos células de distinta estirpe para dar lugar a otra de características mixtas.
>
> (Genética) f. unión entre dos individuos con fenotipos o genotipos distintos, o bien, procedentes de dos poblaciones o especies diferentes. En biología molecular, el emparejamiento específico entre cadenas complementarias de DNA o ácido ribonucleico (RNA).

En el ejemplo 3, se presentan los sustantivos de la clase *body* y se refieren en su mayoría a partes de cuerpo.

3. alelo, arteria, cromosoma, gen, hormona, membrana, mucosa, médula, región, vértebra, vías, zona, área.

Sin embargo, obsérvese que también existen algunas sustancias como hormona que podrían estar bajo la clase *substance* como se ve en las definiciones del diccionario Espasa.

> (Fisiología) f. Sustancia química secretada por las glándulas endocrinas, que alcanza el órgano diana a través de la sangre.

De igual modo, algunos ejemplos como alelo, cromosoma y gen pueden estar bajo la clase *cell* como se deduce de las definiciones del diccionario Espasa de medicina.

> (Gen.) m. cada una de las formas en que puede presentarse un gen en un determinado locus (v.).
>
> (Gen.) m. cada una de las pequeñas formaciones estructurales en forma de bastoncillo en que se divide la cromatina del núcleo celular en la mitosis.
>
> (Biol.) secuencia de ADN que constituye la unidad funcional para la transmisión de los caracteres hereditarios.

Como sucede con esta clase en inglés, existen algunos casos que podrían considerarse como hiperónimos ya que no son exactamente partes del cuerpo sino palabras o paratérminos que se refieren a lugares más amplios no definidos que podrían albergar a partes del cuerpo como puede ser área, región y zona.

Clase semántica	Frecuencia	Porcentaje
state	42	21
act	24	12
body	22	11
cell	15	7,5
attribute	14	7
event	13	6,5
cognition	11	5,5
artifact	9	4,5
process	9	4,5
substance	9	4,5
person	6	3
animal	5	2,5
relation	5	2,5
time	3	1,5
communication	2	1
group	2	1
location	2	1
phenomenon	2	1
object	1	0,5
quantity	1	0,5
not found	3	1,5

Tabla 2: Clases semánticas de los núcleos en EuroWordNet.

También es importante destacar las clases *cell* (7,5%), *attribute* (7%) y *event* (6,5%), como se ejemplifica en 4, 5 y 6.

4. aceptor, cDNA, cepa.

5. concentración, defecto, deficiencia, desequilibrio, estructura, expresión, fenotipo, función, hipermotilidad, homología, temperatura.

6. alteración, cambio, factor.

Los ejemplos de 4 se refieren a palabras relacionadas con células excepto el aceptor que es una sustancia, como lo muestra la definición del Mosby.

> sustancia o compuesto que se combina con una parte de otra sustancia o compuesto.

Los ejemplos de 5 son sustantivos que reflejan atributos como se observa refleja más claramente el etiquetaje de segundo orden (*Cause, Location, Manner, Physical, Property*). Según la documentación de EWN[66], la clase *attribute* se refiere a sustantivos que denotan cualidad de un objeto o entidad: *EWN Static Situation which applies to a single concrete entity or abstract Situation; e.g. colour, speed, age, length, size, shape, weight.*

Es importante anotar que la clase *act* no es una de las más frecuentes en la modificación pero sí en los núcleos, lo que demuestra el que estas nominalizaciones tiendan a ir en el núcleo y los estados (*state*, 11,33%) y adjetivos generales (*adj.all*, 17%) en la modificación, situación similar se ha descrito en inglés.

[66] Para mayor información sobre las categorías de EuroWordNet se puede consultar el sitio http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl.

En 7 se recogen los ejemplos de unidades léxicas (N, Adj y Adv) que no son términos, pero que pueden ayudar en la identificación de unidades terminológicas o que adquieren carácter terminológico dentro de un contexto o área determinados y que pueden denominarse paratérminos.

7. activo, adulto, agudo, alto, anterior, baja, central, centro, corto, diferente, difuso, doble, dominante, dominio, específico, factor, familiar, fino, fuerte, general, generalizado, horizontal, interno, masa, normal, región, sensible, simple, sistema, zona.

7.4.2 Análisis de las clases semánticas de la modificación según EuroWordNet

Para mantener la consistencia con las categorías de WordNet 2.1, se emplearán las mismas clases que EWN emplea en su primer nivel pero de la cual no da una definición. Sin embargo, debe ser igual a la de WordNet 2.1. Por el contrario, tiene clases más subespecificadas de las cuales proporciona información. Así, se agruparán todos los ejemplos pero en algunos casos se empleará la segunda etiqueta para indicar con más precisión la clase semántica a la que pertenece la palabra. Por ejemplo, en el caso de "crónico" que aparece en el primer nivel como *all* y en el segundo como *DiseaseOrSyndrome+*. Así, aunque en WordNet pueda considerarse un adjetivo general, en EWN hereda las características de enfermedad que en WordNet están ligadas a la clase *attribute*.

En la tabla 3, se presentan las clases semánticas de la modificación en EuroWordNet.

La clase semántica más frecuente es *body* con un 24,88% de todas las ocurrencias, seguida por la clase *adj.all* con un 17% y luego la clase *substance* con un 14,04% del total de las ocurrencias y representan un 55,92% de toda la modificación.

Clase semántica	Frecuencia	Porcentaje
body	101	24,88
adj.all	69	17
substance	57	14,04
state	46	11,33
cell	41	10,1
adj.pert	16	3,94
act	15	3,69
process	10	2,46
attribute	8	1,97
person	7	1,72
artifact	6	1,48
cognition	6	1,48
animal	4	0,99
phenomenon	4	0,99
time	4	0,99
event	2	0,49
group	2	0,49
object	2	0,49
communication	1	0,25
location	1	0,25
quantity	1	0,25
not found	3	0,74

Tabla 3: Clases semánticas de la modificación en EuroWordNet.

En 8, se pueden ver los ejemplos de la clase *body* que denotan partes del cuerpo o adjetivos denominales como alélico, craneano, epitelial, etc. Igualmente, hay algunos ejemplos que no se refieren directamente a partes del cuerpo sino a rasgos o aspectos anatómicos como morfológico, resistente, liso, sexual.

8. alelo, alélico, articulación, calloso, cerebro, columna, colágeno, corazón, coronario, craneano, cromosoma, cromosómico, cuello, cuerpo, cutáneo, dorsolumbar, embrionario, epidérmico, epitelial, espinal, fetal, fibroblasto, gen, génico, hepático, hipofisarias, hormona, humano, intestinal, intestino, intravascular, leucocito, liso, macrófago, mama, mano, morfológico, muscular, médula, músculo, nervios, nervioso, neural, periventricular, renal, resistentes, seminífero, sexual, supresor, túbulo, urinario, uterino, vertebral, vesicales, visual, vía, ósea.

Del mismo modo que en inglés se observa que algunos ejemplos de la clase *body* pertenecer directamente a la clase *substance*, como en los ejemplos de colágeno, hormona y alelo (alélico) aunque esta clase aparece en ambos casos en el último nivel de etiquetaje.

Sin embargo, puede observarse que hay algunos ejemplos que pueden clasificarse dentro de la clase *cell*, como cromosoma, cromosómico, leucocito, macrófago, fibroblasto, gen y génico.

En 9, se enseñan los ejemplos de la clase *all* que reúne una serie de subclases en EWN (*SubjectiveAssessmentAttribute, familyRelation, AnatomicalStructure, DiseaseOrSyndrome, equal, located, ShapeAttribute, Process, BiologicallyActiveSubstance, capability, forall, FrontFn, Horizontal, MultiplicationFn, Planning, PositionalAttribute, Radiating, RadiatingLight, RegulatoryProcess, SentientAgent, SoundAttribute*). De estas subclases podemos distinguir *SubjectiveAssessmentAttribute* con 24 ocurrencias (agudo, alto, bajo, corto, delgado, diseminado, generalizado, largo, limitado, normal, primario, proteico, recurrente, simple, superficial), *familyRelation* con 9 ocurrencias (genético, hereditario), *AnatomicalStructure* con 6 ocurrencias (dominante, recesivo), *DiseaseOrSyndrome* con 4 ocurrencias (crónico, degenerativo, retardado), *equal* con 4 ocurrencias (diferente, único).

9. activo, agudo, alto, anterior, apical, bajo, central, consecutivo, corto, crónico, cíclico, degenerativo, delgado, diferente, difuso, diseminado, distal, doble, dominante, falciforme, fino, fuerte, fértil, general, generalizado, genético, hereditario, horizontal, interna, largo, limitado, normal, primario, programado, proteico, radiactivo, recesiva, recurrente, regulador, retardado, sensible, simple, superficial, único.

En 10, pueden verse los ejemplos de la clase *substance* que muestra mucha consistencia y no es necesario recurrir a los otros niveles de etiquetaje para su descripción.

10. agarosa, aminoácido, amoxicilina, bioquímico, bromuro, CFTR, cistina, distrofina, DNA, enzima, etidio, fenilalanina, ferritina, gel, gentamicina, GMP, guanosina, hidroxilasa, hidrógeno, hierro, metilmercurio, mezlocilina, miofosforilasa, mioglobina, molecular, nitrogenado, nucleótido, nítrico, orina, PCR, prolactina, proteína, purina, quimioquina, SHV-1, sérico, tóxica, Vav, óxido.

A diferencia del corpus del inglés no se observa un predominio en las clases *substance* y *body* si bien esta última es una de las tres más frecuentes en ambos casos.

Igualmente, cabe destacar también las categorías *state* con un 11,33%, *cell* con un 10,1% y *pert* con un 3,94%, como se puede en los ejemplos 11, 12 y 13.

11. acondroplasia, adenomatoso, adipogénico, autosómico, botonosa, desnaturalización, enfermedad, fibrosis, fractura, grado, hepatitis, inmunodeficiencia, insulinodependiente, libertad, mellitus, mieloide, pigmentario, potencial, quístico, retinoblastoma, retinosis, riesgo, serotoninérgico, trófico, tumoral, tumor.

Como puede observarse en los ejemplos, predomina la subclase *DiseaseOrSyndrome* con 27 ocurrencias (candidiasis, cáncer, diabetes, fibrosis, insuficiencia, etc.), seguida muy de lejos por la clase *attribute* con 5 ocurrencias (grado). En general, denota enfermedades (hepatitis) o estados de enfermedades (riesgo, grado, potencial).

12. adipocito, autosoma, celular, codificante, célula, fijador, genoma, gramnegativo, grampositivo, intracelular, locus, minisatélite, mitocondrial, mutador, plásmido, portador, promotor, Rad51, supresor, vírico, YAC.

En 13, pueden verse los ejemplos de la clase *cell*, la cual presenta consistencia como clase semántica mientras que los ejemplos de 12 de la clase *pert* presentan una gama variada de adjetivos relacionales como lo reflejan las

subclases a las que pertenecen (*FieldOfStudy, Number, Organism, Organization, Planning, ShapeAttribute, SubjectiveAssessmentAttribute, WaterArea, familyRelation*).

13. biológico, cilíndrico, clínico, específico, familiar, logística, mediterráneo, numéricas, transicional.

Es importante resaltar que no se le asignó ninguna clase semántica en EuroWordNet a más del 0,67% de los casos (6) y a 100 registros de 892 (11,2%) no se le encontró la clase semántica directamente pero se le asignó con la ayuda de los diccionarios de referencia de medicina. Así, el total de ocurrencias que no se encuentran en EuroWordNet se eleva a un 11,88%, lo que puede afectar el etiquetaje automático substancialmente e incidir notablemente en los resultados si no se etiqueta manualmente. A estas unidades se les asignó las clases de EuroWordNet pero señalando dichas unidades como *not found* como es el caso de *poliposis* a la que se le ha asignado finalmente la clase *state* pero que en la base de datos se ha marcado también como *not found.state*. A diferencia del corpus del inglés, en español la mayoría de unidades que no se encontraron no son siglas salvo los ejemplos de 14 y muchas de ellas se encuentran en un diccionario general de medicina como el Mosby (50 casos de 74), como algunos de los casos de 15.

14. CFTR, GMP, PCR, Perls, cDNA, Rad51, SHV-1, Vav, YACS

15. aceptor, agarosa, agenesia, alineamiento, amoxicilina, artrosis, autosoma, autosómica, betalactamasa, cepa, distrofina, dorsolumbar, ferritina, fetopatía, fijador, genoma, gramnegativa, grampositiva, guanosina, inmunodeficiencia.

7.4.2.1 Clases semánticas de acuerdo con la posición dentro de la modificación

Si se observan las clases semánticas de la modificación de acuerdo con la posición dentro del sintagma veremos que no existen diferencias importantes y que el principio de polaridad observado en inglés no es tan claro en español ya

que los *adj.all* son la segunda clase más frecuente en todas las posiciones. Además, las clases como *attribute* y *state* no presentan una frecuencia alta en las diferentes posiciones.

Así que las características menos estables o más subjetivas tienden a ir más alejadas de núcleo del sintagma en español como lo planteado en capítulo anterior para el inglés.

De entrada, la tabla 4 muestra que los datos en las primeras 5 clases y el orden varían muy poco. El predominio de estas clases semánticas se debe en principio al área temática y principalmente las clases *body*, *susbtance*, *cell* y, de cierto modo la clase *state* con la subclase de enfermedades.

Primera posición	Porcentaje	Segunda posición	Porcentaje	Tercera posición	Porcentaje
body	22,5	body	25,37	body	83,33
all	16	all	17,91	all	16,67
substance	12,5	substance	15,92		
state	12	cell	10,95		
cell	9,5	state	10,95		
act	5,5	pert	5,97		
process	4	person	2,99		
attribute	3	act	1,99		
cognition	2	artifact	1,49		
pert	2	animal	1		
time	2	attribute	1		
artifact	1,5	cognition	1		
animal	1	phenomenon	1		
event	1	process	1		
group	1	object	0,5		
phenomenon	1	not found	1		
communication	0,5				
location	0,5				
object	0,5				
person	0,5				
quantity	0,5				
not found	1				

Tabla 4: Clases semánticas por posiciones en la modificación.

Sin embargo, si se suman los adjetivos generales etiquetados como *adj.all* en cada posición más los adjetivos relacionales *adj.pert*, se observa que el

promedio de adjetivos aumenta a medida que la modificación se aleja del núcleo a la izquierda.

Hay 56 palabras (de 200) en la primera posición con adjetivos (28% de posibilidades de ser adjetivo), de los cuales 24 son *adj.pert* y 32 *adj.all,* como se ven en los ejemplos de 16.

16. activo, agudo, alto, anterior, bajo, biológico, celular, clínico, consecutivo, coronario, corto, cutáneo, degenerativo, diferente, distal, doble, embrionario, epitelial, fetal, fuerte, general, genético, horizontal, interno, largo, molecular, morfológico, muscular, nervioso, normal, numérico, primario, proteico, renal, resistente, sensible, tóxica, urinario, visual, ósea.

Hay 88 palabras (de 200) en la segunda posición con adjetivos (44% de posibilidades de ser adjetivo), de los cuales 46 son *adj.pert* y 42 *adj.all*, como se ilustra en los ejemplos de 17.

17. aguda, alélico, bioquímico, calloso, celular, central, cilíndrico, cromosómico, crónico, cíclico, delgado, difuso, diseminado, dominante, epidérmico, específico, falciforme, familiar, fino, fértil, generalizado, genético, génico, hepático, hereditario, humano, interna, intestinal, intracelular, limitado, liso, logística, mediterráneo, nervioso, neural, nítrico, programado, radiactivo, recesivo, recurrente, retardado, sexual, simple, superficial, transicional, uterino, vertebral, vírico, óseo, único.

Hay 7 palabras (de 6 en total) en la tercera posición con adjetivos (85% de posibilidades de ser adjetivo), de los cuales 5 son *adj.pert* y 1 es *adj.all,* como se ve en los ejemplos de 17.

17. epitelial, espinal, muscular, regulador, renal, seminífero.

Así el promedio de adjetivos en cada posición tiende a aumentar de izquierda a derecha, es decir, que a medida que el modificador se aleja del núcleo existen más probabilidades de ser adjetivo.

7.4.2.2 Clases semánticas de acuerdo con la categoría léxica

Si se observan las clases semánticas de la modificación de acuerdo con categoría léxica (N, Adj y PP) puede observarse que hay algunas diferencias en las clases semánticas que predominan en cada categoría no del modo tan notable como en inglés.

En los sustantivos predomina la clase *substance* con un 21,67% mientras que en los adjetivos predomina la clase *all* con un 31,98% y en los participios la clase *all* con un 100%. Luego, sigue la clase *body* (20,69%) en los sustantivos y la clase *body* (29,95%) en los adjetivos.

Sustantivos	Porcentaje	Adjetivos	Porcentaje	Participios	Porcentaje
substance	21,67	all	31,98	all	100
body	20,69	body	29,95		
state	11,33	state	11,68		
cell	9,36	cell	11,17		
act	7,39	pert	8,12		
process	4,93	substance	6,6		
attribute	3,94	not found	0,51		
person	3,45				
artifact	2,96				
cognition	2,96				
animal	1,97				
phenomenon	1,97				
time	1,97				
event	0,99				
group	0,99				
object	0,99				
communication	0,49				
location	0,49				
quantity	0,49				
not found	0,99				

Tabla 5: Clases semánticas por categoría léxica en la modificación.

Así, hay 44 registros de sustantivos con *substance* (21,67%), como se muestra en los ejemplos de 19.

19. agarosa, aminoácido, amoxicilina, anticuerpo, betalactamasa, bromuro, CFTR, cistina, cristal, distrofina, DNA, enzima, etidio, fenilalanina, ferritina, gel, gentamicina, GMP, guanosina, hidrógeno, hierro, lipoproteína, metilmercurio, mezlocilina, miofosforilasa, mioglobina, nucleótido, orina, PCR, prolactina, proteína, purina, quimioquina, SHV-1, sulfonilurea, Vav, óxido.

Sin embargo, sólo hay 13 palabras (6,6%) como adjetivos con la clase *substance* como se observa en 20. En cambio, hay 63 (31,98%) de adjetivos generales (*adj.all*).

20. bioquímico, hidroxilasa, molecular, nitrogenada, nítrico, sérica, tóxica.

7.4.3 Patrones semánticos obtenidos de EuroWordNet

En la tabla 6, se presentan los patrones semánticos se han obtenido de EuroWordNet. Se presentan todos los que tienen +2 de frecuencia (40 patrones). Para una muestra de 200 sintagmas seleccionados existen 145 patrones diferentes con una media de 0,72, es decir, menos de un patrón por cada sintagma, lo que muestra una gran variabilidad y pocas posibilidades de generalización. De estos 145 patrones, hay 105 patrones de una sola ocurrencia. Por tanto, puede verse que existen muy pocas regularidades en los patrones semánticos salvo por los primeros casos.

A continuación, se presentan los patrones semánticos más frecuentes con los patrones sintácticos que ellos representan.

Patrón semántico	Frecuencia	Porcentaje
state body all	7	3,5
cell body body	4	2
state body body	4	2
act substance body	3	1,5
act substance substance	3	1,5
attribute substance substance	3	1,5
body cell state	3	1,5
event process body	3	1,5
state body person	3	1,5
state cell all	3	1,5
state substance substance	3	1,5
act act all	2	1
act process pert	2	1
all attribute act	2	1
animal all body	2	1
artifact cell all	2	1
artifact cell pert	2	1
attribute substance body	2	1
body all cell	2	1
body body body	2	1
body cell body	2	1
cell body substance	2	1
cell state state	2	1
cognition state all	2	1
event phenomenon state	2	1
person state cell	2	1
person state state	2	1
process body all	2	1
process state all	2	1
process substance substance	2	1
relation state body	2	1
state act body	2	1
state body pert	2	1
state cell pert	2	1
state cognition cell	2	1
state state all	2	1
state state pert	2	1
substance animal cell	2	1
time all state	2	1

Tabla 6: Patrones semánticos obtenidos con EuroWordNet.

El patrón semántico *state body all* es el más frecuente con 7 ocurrencias y se correlaciona sintácticamente con el patrón N Adj Adj/PP[67] en 5 ocurrencias y N Prep N Adj con 2 ocurrencias, como se ve en los ejemplos de 21.

21. candidiasis cutánea generalizada, atrofias musculares difusas, tumores vesicales superficiales, insuficiencia renal aguda, insuficiencia renal crónica; cáncer de mama hereditario.

El patrón *cell body body* es también el más frecuente con 4 ocurrencias y se correlaciona sintácticamente con el patrón N Prep N Adj en 3 ocurrencias, como se aprecia en los ejemplos de 22.

22. células de la médula ósea, células de músculo liso.

Luego continua el patrón *state body body* con 4 ocurrencias, y se correlaciona sintácticamente con el patrón N Prep N Adj en todas las ocurrencias, como se indica en los ejemplos de 23.

23. agenesia de cuerpo calloso, carcinoma de cuello uterino, enfermedades de la columna vertebral.

El patrón *act substance body* tiene 3 ocurrencias, 2 de ellas tiene como patrón sintáctico N Prep N Adj, como se ve en los ejemplos de 24.

24. índice de hierro hepático.

El patrón *act substance substance* tiene 3 ocurrencias y todas se correlacionan sintácticamente con el patrón N Prep N Prep N, como se enseña en los ejemplos de 25.

---

[67] Aunque para los aspectos estadísticos se han separado las clases Adj y PP, es conveniente juntarlas para obtener más regularidades. Cuando exista este caso, se indicará N Adj Adj/PP que debe interpretarse como dos patrones en realidad N Adj Adj y N Adj PP.

25. secuencia de aminoácidos de CFTR, secuencia de aminoácidos de SHV-1, tinción con bromuro de etidio.

Finalmente, se destaca el patrón *attribute substance substance* con 3 ocurrencias que tiene el mismo patrón sintáctico, N Prep N Adj, como se observa en los ejemplos de 26.

25. funciones de l óxido nítrico, concentración de ferritina sérica, deficiencia de fenilalanina hidroxilasa.

Puede verse que a pesar de que existen pocas regularidades en los patrones semánticos, existen al interior de cada uno correlaciones con los patrones sintácticos regulares y algunos patrones se perfilan como predominantes en el uso de estas estructuras semánticas: N Prep N Adj y N Adj Adj/PP.

7.4.4 Patrones semánticos en la modificación según EuroWordNet

A pesar de las pocas regularidades que se encuentra en los patrones completos, se observan más regularidades a nivel de la modificación como se observa en la tabla 7.

De los 79 patrones semánticos totales, 40 presentan una sola ocurrencia, 15 con dos ocurrencias, 11 con tres ocurrencias y 2 con cuatro ocurrencias. Además, se encuentran patrones de 5 ocurrencias hasta 13 ocurrencias. Los patrones semánticos de la modificación con más de 5 ocurrencias: *substance substance* con el 6,5% (13 ocurrencias), *body all* con el 6% (12), *body body* con el 6% (12), *all body* con el 4,5% (9), *state all* con el 3,5% (7), *state body* con el 3,5% (7), *substance body* con el 3,5% (7) y *body substance* con el 3% (6). En total representan el 36,5% de toda la modificación. Igualmente, todos los

patrones semánticos presentados en la tabla 7 representan el 80% de toda la modificación.

En el caso de *substance substance* está relacionado sintácticamente con los patrones N Prep N Adj con 6 ocurrencias, N Prep N Prep N con 5 ocurrencias y N Adj Prep N con 2 ocurrencias. Con el patrón *body all*, 9 casos pertenecen al patrón N Adj Adj/PP y 3 casos a N Prep N Adj. En el patrón *body body* hay 8 casos que pertenecen al patrón N Prep N Adj, 3 casos a N Adj Prep N y 1 caso a N Prep N Prep N. El patrón *all body* tiene 7 casos con el patrón N Adj Prep N, 1 con el patrón Adj N Prep N (una variación del anterior) y 1 caso con N Adj Prep N Adj. El patrón *state all* está representado en su totalidad por el patrón N Adj Adj. El patrón *state body* tiene tres patrones: N Prep N Adj con 5 ocurrencias, N Adj Prep N y N Adj Adj con una, respectivamente. El patrón *substance body* tiene también dos patrones N Prep N Adj con 6 ocurrencias y N Adj Prep N con 1. Finalmente, el patrón *body substance* tiene los patrones N Adj Prep N con 4 ocurrencias y N Prep N Prep N con 2.

De los anteriores datos, puede deducirse que los patrones más comunes en la modificación son parte de los patrones sintácticos más frecuentes N Prep N Adj (28) y N Adj Prep N, (17), seguidos de N Prep N Prep N y N Adj Adj; todos ellos son los 4 más frecuentes del corpus de análisis en español y están entre los 5 más frecuentes del corpus lexicográfico del español. Tal y como se comentó en el capítulo 6, puede afirmarse que estas tendencias son generales a este tipo de discurso especializado. Sin embargo, no es posible afirmar esto para otros tipos de discurso en el aspecto semántico ya que éste está muy ligado al área temática y lo más probable es que estos patrones varíen de área en área temática. Aún así, puede aseverarse que existen estructuras semánticas que subyacen a ciertos tipos de sintagmas en cada área temática como se ha visto en este apartado.

Obsérvese que los patrones N Prep N Adj, N Adj Prep N, N Prep N Prep N y N Adj Adj son los que están presentes en las estructuras semánticas más frecuentes y a su vez, son los cuatro patrones más frecuentes tanto en el corpus de análisis como en el corpus lexicográfico. En este sentido, puede afirmarse

que las estructuras más frecuentes tienden a estar correlacionadas semántica y sintácticamente.

Patrón modificación	Frecuencia	Porcentaje
substance substance	13	6,5
body all	12	6
body body	12	6
all body	9	4,5
state all	7	3,5
state body	7	3,5
substance body	7	3,5
body substance	6	3
state state	6	3
all state	5	2,5
cell all	5	2,5
cell body	4	2
cell pert	4	2
act all	3	1,5
act body	3	1,5
act substance	3	1,5
all act	3	1,5
all cell	3	1,5
all substance	3	1,5
body person	3	1,5
body pert	3	1,5
cognition cell	3	1,5
process body	3	1,5
time cell	3	1,5
act cell	2	1
all attribute	2	1
all person	2	1
all phenomenon	2	1
animal cell	2	1
attribute all	2	1
body animal	2	1
body state	2	1
cell state	2	1
phenomenon state	2	1
process pert	2	1
state cell	2	1
state pert	2	1
substance all	2	1
substance state	2	1

Tabla 7: Patrones semánticos en la modificación según EuroWordNet.

A pesar de la frecuencia presentada por los patrones N Prep N Adj, N Adj Prep N, N Prep N Prep N y N Adj Adj en los patrones semánticos, es importante mostrar que estos patrones sintácticos también presentan otros patrones semánticos diferentes de los expuestos.

El patrón N Prep N Adj presenta además del patrón *body body* con 8 ocurrencias, el patrón *substance body* con 6 ocurrencias y el patrón *substance substance* con 6 ocurrencias también. Los patrones *state body* y *state state* tienen 5 ocurrencias y los patrones *cell all* y *cell pert* con 4 ocurrencias. Además, este patrón sintáctico presenta 4 patrones con 3 ocurrencias, 10 patrones con 2 ocurrencias y 16 con 1 sola ocurrencia.

El patrón N Adj Prep N presenta además del patrón *all body* con 14 ocurrencias, el patrón *body substance* con 8 ocurrencias. Los patrones *all substance* y *body body* tienen cada uno 6 ocurrencias. Los patrones *all cell, all state, body animal, cell body, substance state* y *substance substance* tienen 4 ocurrencias. Además, este patrón sintáctico presenta 15 patrones semánticos de 2 ocurrencias.

El patrón N Prep N Prep N presenta además del patrón substance substance con 5 ocurrencias y el patrón *act substance* con 3 ocurrencias. Además, este patrón sintáctico presenta 2 patrones semánticos de 2 ocurrencias y 13 de 1 ocurrencia.

Finalmente, el patrón N Adj Adj presenta además del patrón semántico *body all* con 9 ocurrencias, el patrón *state all* con 7 ocurrencias y el patrón *state pert* con 2 ocurrencias. El resto de patrones (6) son de una ocurrencia. En este sentido, el correspondiente patrón en inglés Adj Adj N tiene la misma variabilidad semántica.

7.5 Recapitulación

En este capítulo, se han presentado los resultados del análisis semántico del corpus de análisis en español en EuroWordNet.

1. Las clases semánticas más frecuentes en el núcleo en EuroWordNet son *state* (21%), *act* (12%), *body* (11%), *cell* (7,5%) y *attribute* (7%). Estas cinco clases semánticas representan el 58,5% de todos los núcleos de la muestra.

2. Las clases semánticas más frecuentes en la modificación en EuroWordNet son *body* (24,88%), *adj.all* (17%), *substance* (14,04%), *state* (11,33%) y *cell* (10,1%). Estas cinco clases semánticas representan el 77,35% de toda la modificación de la muestra. Un aspecto importante en la modificación es que el promedio de adjetivos en cada posición tiende a aumentar de izquierda a derecha, es decir, a medida que el modificador se aleja del núcleo existen más probabilidades de ser adjetivo.

3. Los patrones más frecuentes en EuroWordNet son *state body all* (3,5%), *cell body body* (2%), *state body body* (2%), *act substance body* (1,5%) y *act substance substance* (1,5%). Al igual que en inglés, estos patrones semánticos obtenidos a partir de EuroWordNet tan solo representan el 10,5% de la muestra de análisis.

4. Puede verse que, al igual que en inglés, no es posible obtener muchas generalizaciones en cuanto a los patrones ya que cada patrón semántico no abarca a más del 3,5% de todas las ocurrencias en el mejor de los casos. Sin embargo, los patrones más frecuentes se correlacionan sintácticamente con dos de los patrones superficiales más frecuentes tanto en el corpus de análisis en español como en el lexicográfico: N Prep N Adj y N Adj Adj/PP.

5. Al igual que en inglés, los resultados reflejan lo “esperable” en cuanto a las clases semánticas puesto que el área temática de este estudio, el genoma, tiene involucradas estas clases. Aunque su aporte puede trasladarse a otras disciplinas del lenguaje como se indicó en §6.5.

6. Aun así, es importante señalar las limitaciones de EuroWordNet. En primer lugar, existe un 11,88% de palabras que no se encuentran en EuroWordNet. Esto puede o que puede afectar el etiquetaje automático substancialmente e incidir notablemente en los resultados si no se etiqueta manualmente. Finalmente, debe tenerse en cuenta las inconsistencias en el etiquetaje de ambos sistemas que se han explicado para poder obtener más regularidades en los resultados.

7. Si bien WordNet y EuroWordNet están relacionadas y se han desarrollado de manera independiente, ambas tienen un silencio considerable en el etiquetaje de las muestras (7,4% vs. 11,88%). Igualmente, en ambos sistemas no se pueden obtener muchas regularidades en los patrones ya que en ambos casos el patrón más frecuente no representa a más del 3,5% de la muestra. En cuanto a las diferencias del etiquetaje en inglés y español en el caso de las inconsistencias en el etiquetaje, EuroWordNet no presenta este tipo de problemas ya que en algún punto del etiquetaje presenta la clase que soluciona la inconsistencia, como se explicó en §7.4.2.

8. Descripción y análisis de los sintagmas nominales en el corpus paralelo

8.1. Introducción

En los capítulos anteriores, se han descrito cuantitativa y cualitativamente los sintagmas nominales extensos especializados en inglés y español para dar cuenta del fenómeno en ambas lenguas. En este capítulo, se pretende observar el comportamiento de estos en un corpus paralelo para retomar el segundo objetivo de esta tesis: demostrar que existen regularidades en tratamiento de los sintagmas nominales del inglés al español para propósitos de traducción, principalmente. Pero, ¿puede estar este tipo de sintagmas interferidos por la sintaxis del inglés? Así, se pretende demostrar que las tendencias en la traducción de este tipo de sintagmas no se deben a interferencias sintácticas del inglés. Para ello, se comparan los resultados del corpus paralelo con los encontrados en el diccionario Mosby de medicina y, después las soluciones proporcionadas por los traductores en los textos y las encontradas en el diccionario Mosby se contrastaran con los patrones obtenidos en el *Corpus Tècnic* del Iula y la consulta hecha al CREA de la RAE. De este modo, se podrá observar si la frecuencia de las soluciones en español es similar a la frecuencia de dichas estructuras en español y establecer si los traductores emplean intuitivamente estructuras del español.

8.2. Recolección del corpus paralelo y extracción de los datos

Para este capítulo se compiló y procesó un corpus paralelo inglés-español que se empleó como corpus de análisis (ver §3.4.1).

Además, se han empleado tres corpus diferentes de contraste: un corpus especializado en español del *Corpus Tècnic* del IULA, el corpus CREA de la RAE

y un corpus lexicográfico en español. Uno de los diccionarios del corpus lexicográfico, el diccionario Mosby de medicina, se empleó para validar los patrones más frecuentes en inglés y sus patrones equivalentes en español encontrados en el corpus paralelo.

En cuanto al corpus CREA, se solicitó una consulta a la RAE con los mismos patrones en español que se usaron en la primera extracción de datos para el *Corpus Tècnic* del IULA. Dicha consulta se realizó sobre un corpus de 5.397 documentos y 143.440.437 tokens.

Con el fin de estudiar la traducción de la premodificación del inglés al español, se recopiló un corpus de 66.534 palabras de 21 textos ingleses. Todos los textos son artículos de investigación que siguen la estructura *Introducción, Materiales y Métodos, Resultados y Discusión* (*IMMRAD*) de la revista médica *The Lancet* como ya se explicó en §3.1.4. Puesto que esta revista se tradujo íntegramente al español hasta 1999, todos los textos se seleccionaron de 1997 a 1998.

Todos los sintagmas se extrajeron manualmente excepto aquellos con posesivo, núcleos coordinados con *and* y *or*. No se incluyeron clases léxicas cerradas en el sintagma excepto aquellas que forman parte de la premodificación, según se indica en los ejemplos 1 a 4.

1. time-to-treatment subgroups
2. on-going clinical trials
3. quality-of-life analysis
4. time-to-definitive-quality-of-life deterioration

Para el caso del español, se excluyeron los determinantes que aparecen al inicio del sintagma nominal y se dejaron las preposiciones y determinantes dentro del sintagma nominal. Sin embargo, se eliminó el determinante con el fin de obtener los patrones y la longitud en español. En este sentido, no se tuvo en

cuenta la preposición. Como se indica en 5, ambos sintagmas nominales tienen el mismo patrón, pero el segundo tiene el determinante (la).

5.	bajo riesgo de ictus documentado	Adj N Prep N Adj
	posterior disminución de la activación inmunitaria	Adj N Prep N Adj

Se ha contado el número de tokens de cada sintagma nominal y se realizaron los cálculos estadísticos básicos (p. ej., longitud, frecuencia de patrones, selección de la muestra, etc.). Cuando fue necesario, se corrigió manualmente el etiquetaje morfosintáctico.

8.3. Resultados

En inglés, se extrajeron 1.724 sintagmas (1.649 sintagmas sin repetición) de la totalidad del corpus. Para los análisis cuantitativos, se tendrán en cuenta los sintagmas que no están repetidos.

8.3.1. Longitud y frecuencia de los sintagmas nominales

Tal y como se afirmó en la introducción, la longitud de un sintagma es inversamente proporcional a su frecuencia. Por tanto, cuanto más extensa sea la premodificación, más especializado e inestable puede ser el sintagma. Este hecho podría llevar a la estabilización y posible lexicalización de un sintagma, lo que confirmaría la idea de que existe una relación directa entre longitud, grado de especialización y estabilización sintáctica.

La longitud de los sintagmas varía de 3 tokens (siendo uno el núcleo) hasta ocho (siendo uno el núcleo) según lo indican los ejemplos 6 y 7.

6.	orthotopic liver transplantation

7. human acute lymphoblastic leukemia CCRF-CEM cDNA library

8.3.1.1. Distribución de longitud entre sintagmas nominales

En el corpus, los sintagmas de 3 tokens son los más frecuentes (1.064 ocurrencias correspondientes al 64,3%) según se observa en la tabla 1. Por el contrario, los sintagmas de 7 o más tokens son los menos frecuentes (13 ocurrencias que corresponden al 1,1%).

8.3.1.2. Distribución según el número de tokens

Los 1.649 sintagmas están distribuidos según el número de tokens como se indica a continuación.

Longitud	Ocurrencias	Porcentaje
3	1.064	64,3
4	401	24,3
5	132	8
6	38	2,3
7	14	1,1
Total	1.649	100%

Tabla 1: Número de ocurrencias y porcentaje distribuidos por longitud.

Estos resultados confirman las observaciones efectuadas en un estudio previo (Quiroz 2005) y Cartagena (1998) en la que la estabilización de sintagmas oscila entre tres y cuatro tokens. En el corpus de este estudio, corresponden al 88,6%. Si se comparan estos datos con los estudios presentados en §1.3, la longitud más frecuente oscila entre tres y cuatro a pesar de los diferentes métodos y corpus empleados. En este estudio solo el 11,4% de los sintagmas representan al resto de las ocurrencias. Sin embargo, al revisar los datos, puede observarse que la mayoría de ellos tienden a ser términos muy complejos. Aunque es posible encontrar menos sintagmas con palabras especializadas en sintagmas de 3 tokens. Desde el punto de vista terminológico, esto es muy relevante puesto que la mayoría de estos sintagmas extensos no

están incluidos en diccionarios según lo demuestra Burgos (2006: 350-353). Si se toma en cuenta el tiempo empleado por un traductor en las búsquedas terminológicas reportado por Fähndrich (2005: 239) - un promedio de 40% del tiempo que toma la traducción- esta clase de sintagma causaría problemas importantes incluso para un traductor experimentado. A su vez, aquellos que son los más difíciles de resolver presentan más variación gramatical en los patrones superficiales según se presenta en §3.3.

8.3.2. Categoría léxica predominante en la premodificación del corpus paralelo

Como se comentó antes, las gramáticas inglesas más importantes (Biber *et al* 1999: 589) y literatura de inglés propósitos específicos privilegian al adjetivo y no al sustantivo como la categoría léxica más común en la premodificación. No obstante, este estudio también confirma las observaciones hechas antes (Quiroz, 2005) en el sentido de que los sustantivos son más frecuentes dentro de la premodificación que los adjetivos (1.881 sustantivos premodificadores) en el discurso especializado. Esto podría explicarse por el hecho de que el discurso especializado usa la nominalización como estrategia discursiva para expresar impersonalización y objetividad. Puesto que los objetos, procesos y acciones se emplean para representar el conocimiento de un área temática, la premodificación es una manera efectiva de acumular sustantivos y comprimir oraciones.

Categoría	Ocurrencias	Porcentaje
N (sin núcleos)	1.881	32,56
Adj	1.865	32,23
PP	210	3,63
Adv	98	1,69
PPi	40	0,69
Prep	31	0,54
Conj	11	0,19
V	2	0,03

Tabla 2: Ocurrencias y porcentajes de las categorías léxicas.

8.3.3. Frecuencia de patrones en inglés

Los patrones se pueden analizar según su frecuencia en todos los sintagmas al igual que su distribución en longitud. Así los 157 patrones superficiales representan a 1.649 sintagmas, pero sólo los primeros 20 patrones corresponden a la mayoría de las ocurrencias (1.359 sintagmas nominales correspondientes al 82,43%).

Longitud	Patrón	Ejemplo	Ocurrencias	%
3	Adj N N	acute heart failure	359	21,77
3	Adj Adj N	chronic active hepatitis	271	16,43
3	N N N	agarose gel electrophoresis	196	11,89
3	N Adj N	brain natriuretic peptide	74	4,49
4	Adj N N N	abnormal liver function tests	70	4,24
4	N N N N	dihydropyridine calcium channel blocker	59	3,58
4	Adj Adj N N	actual systolic blood pressure	55	3,34
3	PP Adj N	dislodged biliary stent	46	2,79
4	Adj Adj Adj N	global left ventricular dysfunction	38	2,3
3	PP N N	inactivated influenza vaccine	30	1,82
4	N Adj N N	insulin-dependent diabetes Mellitus	26	1,58
4	Adj N Adj N	epidural fibre optic device	24	1,46
3	Adv Adj N	unexpectedly high response	21	1,27
4	N N Adj N	sodium intake dietary recall	19	1,15
3	N PP N	mite-induced rhinoconjunctivitis	16	0,97
3	Adv PP N	serologically proven influenza	13	0,79
5	Adj N N N N	higher baseline CD4 lymphocyte counts	11	0,67
5	N N N N N	Kaplan-Meier survival function estimates	11	0,67
5	Adj Adj N N N	multiple single strand DNA breaks	10	0,61
5	Adj Adj Adj Adj N	symptomatic left ventricular systolic dysfunction	10	0,61

Tabla 3: Los 20 patrones más frecuentes en el corpus.

Todos estos patrones presentan más de 10 ocurrencias y su longitud oscila entre tres y cinco tokens. Hay muchos patrones de tres tokens que representan 1.026 ocurrencias, siete patrones con cuatro tokens que representan 291 ocurrencias, y cuatro patrones de cinco tokens que representan 42 ocurrencias. Los otros 137 patrones que representan solamente 290 ocurrencias (17,57%) demuestran una muy alta variabilidad sintáctica, lo cual

no permite hacer mayores generalizaciones (casi 1 patrón por cada 3 sintagmas). Estos datos indican que la longitud está directamente relacionada con la estabilización de tales estructuras y que una variabilidad mayor está directamente relacionada con una mayor premodificación más extensa.

En términos generales, los patrones más frecuentes son Adj N N, Adj Adj N, y N N N con 826 ocurrencias que corresponden al 50,09% de las ocurrencias, según se presenta en la tabla 3.

Longitud	Patrones	Porcentaje	Ocurrencias
3	21	64,3	1.064
4	43	24,3	401
5	53	8	132
6	27	2,3	38
7	13	1,1	14
Total	157	100%	1.649

Tabla 4. Distribución de patrones por longitud en todos los sintagmas nominales.

La distribución de patrones por número de tokens de la tabla 4 sugiere que los patrones más frecuentes son patrones de 3 tokens (media = 50,6 patrones) y los patrones menos frecuentes son los de 7 tokens (media = 0,9 patrones). Eso significa que la variabilidad de los patrones en los patrones menos extensos es baja comparada con la variabilidad tan alta que aparece en los más extensos. Por tanto, se encontrarían más patrones de 6 ó 7 tokens directamente relacionados con la cantidad de sintagmas nominales (casi un patrón por sintagma). Es importante decir que con el fin de corroborar esto, es necesario realizar procedimientos similares en corpus especializados más grandes con textos del tipo IMMRAD como el que se estudia aquí. Por tanto, las estadísticas básicas que se han discutido aquí deben leerse con cautela.

8.3.4. Frecuencia de patrones por longitud

A continuación se analizan los patrones más comunes distribuidos por longitud y sus respectivos ejemplos[68] en las tablas 5, 6, 7 y 8.

Patrón	Ejemplo	Ocurrencias
Adj Adj Adj Adj N N	mean normal left ventricular ejection fraction	3
Adj N Adj N N N	best prevaccination early-morning PEF values	3
Adj N N N N N	chronic hepatitis C virus (HCV) infection	3
N N N N N N	Mean (SD) log plasma renin activity	3

Tabla 5: Los patrones más frecuentes de 6 tokens.

En la tabla 5, se listan los 4 patrones de 6 tokens. En ellos no hay uno que predomine totalmente. De hecho, todos tienen las mismas ocurrencias. En esta longitud, puede observarse *a priori* que los sintagmas tienden a ser formaciones libres. Sin embargo, en los diccionarios se pueden encontrar ejemplos lexicalizados de esta extensión y aún más extensos.

Patrón	Ejemplo	Ocurrencias
Adj Adj N N N	multiple single-strand DNA breaks	10
Adj N N N N	high baseline HCV RNA concentrations	10
N N N N N	Kaplan-Meier survival function estimates	10
Adj Adj Adj Adj N	symptomatic left ventricular systolic dysfunction	10
Adj Adj Adj N N	acute lower respiratory tract infections	8
Adj Adj N Adj N	simple large-scale clinical trial	6
Adj N Adj N N	standard complement-dependent microcytotoxicity assay	6
Adj N N Adj N	atypical liver kidney microsomal antibody	5
N Adj N N N	Cox proportional hazards regression models	5

Tabla 6. Los patrones más frecuentes de 5 tokens.

En la tabla 6, se muestran los patrones más frecuentes de 5 tokens. Al igual que con los patrones de 6 tokens, en los patrones de 5 tokens no existe un

[68]Estos patrones también se tienen en cuenta para la misma distribución de la muestra seleccionada en §8.3.5.

patrón o conjunto de patrones que predominen claramente. Sin embargo, los 5 primeros patrones presentan ya regularidades en la frecuencia si se tiene en cuenta que a mayor extensión menor frecuencia. Obsérvese que los ejemplos de algunos de estos patrones tienden a ser menos libres y presentan más lexicalización *a priori*.

Patrón	Ejemplo	Ocurrencias
Adj N N N	additional combination therapy group	70
N N N N	baseline HCV RNA concentration	57
Adj Adj N N	gastric parietal cell antibody	54
Adj Adj Adj N	global left ventricular dysfunction	37
N Adj N N	insulin-dependent diabetes Mellitus	25
Adj N Adj N	perinuclear antineutrophil cytoplasmic antibody	24
N N Adj N	liver kidney microsomal antibody	19

Tabla 7: Los patrones más frecuentes de 4 tokens.

En las tablas 7 y 8, se presentan los patrones más frecuentes de 4 y 3 tokens, respectivamente. Éstos presentan unas frecuencias mucho más altas e igualmente presentan *a priori* más lexicalización.

Patrón	Ejemplo	Ocurrencias
Adj N N	autoimmune graft dysfunction	359
Adj Adj N	cardiogenic pulmonary oedema	271
N N N	ACE inhibitor monotherapy	196
N Adj N	brain natriuretic peptide	74
PP Adj N	computed axial tomography	46
PP N N	manipulated sodium intake	30
Adv Adj N	clinically relevant improvement	21
N PP N	phosphate buffered solution	16
Adv PP N	persistently raised aminotransferases	13

Tabla 8: Los patrones más frecuentes de 3 tokens.

En este trabajo no se ha hecho ninguna prueba de lexicalización, pues para los propósitos de este trabajo se tienen en cuenta todos los tipos de sintagmas definidos en la metodología. En primer lugar, porque el traductor debe traducirlos todos sean términos o no, al igual que el aprendiz de lenguajes especializados debe interpretarlos para poderlos entender. En segundo lugar,

porque simplemente son un problema de traducción y no necesariamente de terminología.

Según se puede observar en las tablas presentadas antes, los patrones con tres o más tokens corresponden no sólo a la mayoría de patrones sino también a los más productivos. Esta selección de patrones y ejemplos podría ser relevante para traductores o en la enseñanza de la traducción para identificar los patrones más comunes y dar una solución o aprender a traducirlos. También podrían ser de utilidad para la identificación y extracción de terminología o la traducción automática como se explicó en explicó en §4 y §5.

8.3.5. Selección de la muestra

Se calculó una muestra de 320 sintagmas con un error del 5% con el fin de seleccionar los equivalentes al español y comparar las soluciones con los patrones en inglés seleccionados anteriormente. Los 320 sintagmas se distribuyeron proporcionalmente según la longitud y, a la vez, los patrones más representativos en cada longitud, según se muestra en la tabla 9.

Longitud	Muestra	Ocurrencias	Porcentaje	Patrones
3	205	1.064	64,3	21
4	77	401	24,3	43
5	26	132	8	53
6	10	38	2,3	27
7	14	14	1,1	13
Total	332	1.649	100%	157

Tabla 9: Muestra seleccionada según la longitud.

Los patrones seleccionados son los que se presentan en las tablas 5 a 8 según su frecuencia dentro de la misma longitud. Ya que había pocos ejemplos de 7 ó más sintagmas, se tomaron más ejemplos con el fin de observar las soluciones.

A cada patrón se le asignó su respectiva traducción al español (en algunos casos hasta tres traducciones) y los patrones superficiales en español (etiquetaje).

8.3.6. Clasificación de soluciones de acuerdo con la dependencia sintáctica

En la tabla 10, se presenta una clasificación de las soluciones de acuerdo con la dependencia sintáctica en inglés y sus correspondientes en español.

Longitud	Patrón ENG	Dependencia ENG	Patrón ESP	Dependencia ESP	Frec.	Porc.
3	Adj Adj N	[C [B A]]	N Adj Adj	[[A B] C]	38	18,2
3	Adj Adj N	[C [B A]]	N Adj Adj	Ambiguo	3	1,44
3	Adj Adj N	[C [B A]]	Adj N Adj	[[A B] C]	2	0,96
3	Adj Adj N	[C [B A]]	N Adj Adj	[A [B C]]	2	0,96
3	Adj Adj N	[C [B A]]	N Adj Adj Adj	[[[A B] C] D]	2	0,96
3	Adj Adj N	[C [B A]]	N Prep N Adj	[A [B C]]	2	0,96
3	Adj N N	[[C B] A]	N Prep N Adj	[A [B C]]	15	7,18
3	Adj N N	[C [B A]]	N Prep N Adj	[A [B C]]	12	5,74
3	Adj N N	[C [B A]]	N Adj Adj	[[A B] C]	9	4,31
3	Adj N N	[C [B A]]	N Adj Prep N	[[A B] C]	9	4,31
3	Adj N N	[C [B A]]	N Adj	[A B]	5	2,39
3	Adj N N	[C [B A]]	N N Adj	[[A B] C]	5	2,39
3	Adj N N	[[C B] A]	N Adj Adj	Ambiguo	2	0,96
3	Adj N N	[C [B A]]	Adj N Prep N	[[A B] C]	2	0,96
3	Adj N N	[[C B] A]	N Adj Adj	[[A B] C]	1	0,48
3	Adj N N	[[C B] A]	N Prep N Prep N	Ambiguo	1	0,48
3	Adj N N	[C [B A]]	N Prep N Prep N	[[A B] C]	1	0,48
3	Adv Adj N	[[C B] A]	N Adv Adj	[A [B C]]	2	0,96
3	Adv Adj N	[[C B] A]	N Adj Adj	Ambiguo	1	0,48
3	Adv PP N	[[C B] A]	N Adj Adv	[A [B C]]	2	0,96
3	Adv PP N	[[C B] A]	N Adv PP	[A [B C]]	2	0,96
3	N Adj N	[C [B A]]	N Adj Adj	[[A B] C]	5	2,39
3	N Adj N	[C [B A]]	N Prep N Adj	[A [B C]]	2	0,96
3	N Adj N	[C [B A]]	N Prep N Prep N	[A [B C]]	2	0,96
3	N Adj N	[C [B A]]	N Adj Adj	Ambiguo	1	0,48
3	N Adj N	[C [B A]]	N Adj Prep N	[[A B] C]	1	0,48
3	N N N	[[C B] A]	N Prep N Adj	[A [B C]]	7	3,35
3	N N N	[[C B] A]	N Prep N Prep N	[A [B C]]	7	3,35
3	N N N	[[C B] A]	N Prep N N	[A [B C]]	4	1,91
3	N N N	[C [B A]]	N Adj Prep N	[[A B] C]	4	1,91
3	N N N	[C [B A]]	N Prep N Adj	[A [B C]]	4	1,91
3	N N N	[[C B] A]	N Prep N	[[A B] C]	1	0,48

3	N N N	[C [B A]]	N Prep N	[[A B] C]	1	0,48
3	N N N	[C [B A]]	N Prep N Prep N	[A [B C]]	1	0,48
3	N PP N	[C [B A]]	N PP Prep N	[A [B C]]	2	0,96
3	N PP N	[[C B] A]	N Adj Prep N	[[A B] C]	1	0,48
3	PP Adj N	[C [B A]]	N Adj Adj	[[A B] C]	5	2,39
3	PP Adj N	[C [B A]]	N Adj Adj	Ambiguo	1	0,48
3	PP N N	[C [B A]]	N Adj Prep N	[[A B] C]	2	0,96
3	PP N N	[C [B A]]	N Prep N PP	[[A B] C]	2	0,96
4	Adj Adj Adj N	[D [C [B A]]]	N Adj Adj Adj	[A [B C D]]	2	0,96
4	Adj Adj Adj N	[D [C [B A]]]	N Adj Adj Adj	[A [B C D]]	2	0,96
4	Adj Adj N N	[[D [C B]] A]	N Prep N Adj Adj	[A [[B C] D]]	2	0,96
4	Adj Adj N N	[D [C [B A]]]	N Prep N Adj Adj	[A [[B C] D]]	1	0,48
4	Adj N Adj N	[[D C] [B A]]	N Adj Adj Adj	[[[A B] C] D]	2	0,96
4	Adj N Adj N	[[D C] [B A]]	N Adj Prep Adj N	[[A B] [C D]]	2	0,96
4	Adj N Adj N	[D [C [B A]]]	N Adj Adj Adj	[[[A B] C] D]	2	0,96
4	Adj N N N	[[D C] [B A]]	N Prep N Prep N Adj	[[A B] [C D]]	2	0,96
4	Adj N N N	[D [C B] A]]	N Adj Prep N N	[[A B] [C D]]	2	0,96
4	Adj N N N	[D [C B] A]]	N Adj Prep N Prep N	[[A B] [C D]]	2	0,96
4	Adj N N N	[D [C [B A]]]	N Adj Adj Prep N	[[[A B] C] D]	2	0,96
4	Adj N N N	[[D C] [B A]]	N Prep N Prep N Adj	[[[A [B [C D]]]	1	0,48
4	Adj N N N	[D [C B] A]]	N Prep N Prep N Adj	[[A B] [C D]]	1	0,48
4	Adj N N N	[[[D C] B] A]]	N Adj Prep N Prep N	[[A B] [C D]]	1	0,48
4	Adj N N N	[D [C [B A]]]	N Adj Prep N Prep N	[[A B] [C D]]	1	0,48
4	N Adj N N	[D [C B] A]]	N Prep N Adj Prep N	[A [[B C] D]]	2	0,96
4	N Adj N N	[[[D C] B] A]]	N Prep N Adj Prep N	[A [[B C] D]]	1	0,48
4	N N N N	[D [C B] A]]	N Adj Prep N N	[[A B] [C D]]	4	1,91
4	N N N N	[[C B] A]	N Adj Prep N N	[[A B] [C D]]	1	0,48
4	N N N N	[D [C B] A]]	N Prep N Adj	[A [B C]]	1	0,48
4	N N N N	[[[D C] B] A]]	N Prep N Adj	[A [B C]]	1	0,48
5	Adj Adj Adj Adj N	[E [[D C] [B A]]]	N Adj Adj Adj Adj	[A [B C D E]]	3	1,44

Tabla 10. Clasificación de las soluciones de acuerdo con la dependencia sintáctica.

Puede inferirse que entre más extenso sea el sintagma, más posibilidades de traducción tiene en español. Así, los patrones de 3 tokens tienen, en su mayoría, una sola solución y, en cambio, los patrones de 4 y 5 tokens tienen más de una solución. En general, puede decirse que la solución más común para los sintagmas de 3 tokens para la dependencia [C [B A]] son [[A B] C] con 86 ocurrencias y [A [B C]] con 27 ocurrencias y para la dependencia [[C B] A] es [A [B C]] con 39 ocurrencias.

Para los sintagmas de 4 tokens existen varias dependencias en inglés (6 dependencias) y, por tanto, las posibilidades aumentan en español (6 dependencias). La dependencia en inglés que presenta más regularidad es [[D

C] [B A]] con la dependencia en español [[A B] [C D]] con 4 ocurrencias y para la dependencia en inglés [D [C [B A]]] con [[[A B] C] D] y [A [B C D]] ambas con 4 ocurrencias, respectivamente. Para la dependencia [D [C B] A]] en inglés, la dependencia más regular en español es [[A B] [C D]] con 9 ocurrencias. El único patrón de 5 ocurrencias tiene la misma dependencia en ambas lenguas.

Salvo en algunos casos, puede observarse que no hay interpretaciones lineales del sintagma como lo han planteado muchos autores en traducción (ver §2.5). En primer lugar, las dependencias en inglés no son lineales y, por ende, en español pueden existir variaciones en la interpretación por parte del traductor. En segundo lugar, tampoco existen dependencias lineales en español ya que un mismo patrón puede tener dos interpretaciones si el tipo de adjetivo es relacional o calificativo como se ha visto en §5.3.5.

8.3.7. Resultados del corpus paralelo de acuerdo con el patrón en inglés

En la tabla 11, se presentan los resultados de la comparación de los patrones en inglés y los patrones encontrados en español. De los 20 patrones más frecuentes, 13 presentan alguna regularidad en español. Obsérvese que no existen regularidades en español para patrones de 6 o más tokens.

De estos 13 patrones, sólo un patrón de 5 tokens tiene un patrón común en español, 6 patrones de 4 tokens y 6 patrones de 3 tokens tienen uno o más patrones comunes en español. No puede decir que las regularidades en los patrones del inglés al español este ligada a la extensión ya que los datos actuales no permiten establecer este hecho. No obstante, si se tiene en cuenta que la variabilidad sintáctica aumenta de acuerdo con la extensión es muy probable que las regularidades disminuyan con la extensión.

Longitud	Patrón ENG	Patrón ESP	%	Patrón ESP	%	Patrón ESP	%
5	Adj Adj Adj Adj N	N Adj Adj Adj Adj	100				
4	Adj Adj Adj N	N Adj Adj Adj	100				
4	Adj Adj N N	N Prep N Adj Adj	100				
4	Adj N Adj N	N Adj Adj Adj	66,7	N Adj Prep Adj N	33,3		
4	Adj N N N	N Prep N Prep N Adj	33,3	N Adj Prep N Prep N	33,3	N Adj Prep N N	16,7
4	N Adj N N	N Prep N Adj Prep N	100				
4	N N N N	N Adj Prep N N	71,4	N Prep N Adj	28,6		
3	Adj Adj N	N Adj Adj	87,8	Adj N Adj	4,08	N Prep N Adj	4,08
3	Adj N N	N Prep N Adj	43,5	N Adj Adj	19,4	N Adj Prep N	14,5
3	Adv Adj N	N Adv Adj	100				
3	Adv PP N	N Adj Adv	50	N Adv PP	50		
3	N Adj N	N Adj Adj	50	N Adj Prep N	16,7	N Prep N Adj	16,7
3	N N N	N Prep N Adj	37,9	N Prep N Prep N	27,6	N Adj Prep N	13,8

Tabla 11: Patrón en inglés con los diferentes patrones encontrados en español.

Más de la mitad de los patrones (7) en inglés tienen un único patrón en español y 5 patrones tienen una solución que predomina sobre las otras y 5 patrones que tienen 3 soluciones en español. Tan sólo 3 patrones tienen varios patrones en español en los que ninguno de ellos representa el 50% de las ocurrencias. Aún así, siempre hay un patrón que predomina entre ellos.

Así, se muestra que existen regularidades en los datos del corpus de análisis al menos para los patrones más frecuentes del corpus. A continuación, se contrastan estos datos con el diccionario Mosby y, luego con el CREA de la RAE para observar si existen interferencias en las propuestas obtenidas de los traductores.

8.4. Correlación entre el corpus paralelo y el diccionario Mosby

Con el fin de comprobar si los patrones más frecuentes en inglés del corpus paralelo presentaban las mismas tendencias en cuanto a los patrones equivalentes en español, se empleó el diccionario Mosby de medicina para observar si en un diccionario de la misma área temática y construido de forma clásica (no con metodología de corpus) presentaba tendencias similares.

Para ello, se tomaron los 13 patrones más frecuentes del corpus de referencia de acuerdo con la longitud (de 3 a 5 tokens) y los patrones equivalentes en español más frecuentes y se contrastaron contra los datos del diccionario.

En la tabla 12 se presentan los patrones analizados del corpus paralelo con los primeros patrones en español tanto del corpus paralelo como del diccionario Mosby. Puede observarse que los patrones más frecuentes en español en el corpus paralelo son igualmente los más frecuentes en el diccionario Mosby excepto las soluciones obtenidas para el patrón N N N N.

Cuando existe un patrón en español como segunda variante de un patrón en inglés, dicha variante también aparece en el diccionario Mosby como segunda variante.

Aunque no se presentan en la tabla 12, en el diccionario Mosby existen algunas variantes de baja frecuencia en muchos casos que no existen en el corpus paralelo. Esto puede deberse al tamaño del corpus paralelo que no permite obtener más datos en este sentido.

El patrón N N N N es el único en inglés que no concuerda con ningún patrón en español tanto para el corpus paralelo como para el diccionario Mosby.

Los patrones Adj N Adj N, N Adj N y N N N tienen las mismas soluciones tanto en el corpus paralelo como en el diccionario Mosby, pero varían en el orden ya que la solución más frecuente en el corpus paralelo es la segunda del diccionario Mosby y, la segunda solución encontrada en el corpus paralelo es la más frecuente en el diccionario Mosby.

De acuerdo con los datos presentados puede verse que las soluciones dadas por los traductores no están necesariamente interferidas sintácticamente. De hecho, solo un patrón no presenta ninguna regularidad, como ya se indicó y

esto se debe, en parte, a que la estructura de este patrón, N N N N, presenta múltiples dependencias en inglés: [[N N] [N N]]], [[N N] [N [N]]]], [N [N [N N]]] y [N [[N N] N]]] ya que no hay elementos conectores que permitan explicitar las relaciones entre los diferentes elementos. Así, las posibilidades en español u otra lengua romance son muy variadas aunque esencialmente la explicitación de relación en español hace que los complementos sean de tipo preposicional, como se aprecia en los patrones de la tabla 13.

Corpus ENG	Mosby ESP	%	Mosby ESP	%	Patrón ESP	%	Patrón ESP	%
Adj Adj Adj Adj N	N Adj Adj Adj Adj	100			N Adj Adj Adj Adj	100		
Adj Adj Adj N	N Adj Adj Adj	51,6	N Adj N Adj	12,9	N Adj Adj Adj	100		
Adj Adj N N	N Prep N Adj Adj	33,3	N Adj Adj Adj	15,9	N Prep N Adj Adj	100		
Adj N Adj N	N Adj Prep N Adj	40	N Adj Adj Adj	20	N Adj Adj Adj	66,7	N Adj Prep Adj N	33,3
Adj N N N	N Prep N Prep N Adj	28	N Prep N Adj Adj	16	N Prep N Prep N Adj	33,3	N Adj Prep N Prep N	33,3
N Adj N N	N Prep N Adj Prep N	18,2	N Adj Adj Adj	18,2	N Prep N Adj Prep N	100		
N N N N	N Prep N Prep N Prep N	23,1	N Prep N Prep N	14,3	N Adj Prep N N	71,4	N Prep N Adj	28,6
Adj Adj N	N Adj Adj	72,3	N Adj	4,06	N Adj Adj	87,8	Adj N Adj	4,08
Adj N N	N Prep N Adj	46,3	N Adj Adj	18	N Prep N Adj	43,5	N Adj Adj	19,4
Adv Adj N	N Adv Adj	80	N Adv PP	20	N Adv Adj	100		
Adv PP N		0			N Adj Adv	50	N Adv PP	50
N Adj N	N Adj Prep N	29,1	N Adj Adj	16,4	N Adj Adj	50	N Adj Prep N	16,7
N N N	N Prep N Prep N	35	N Prep N Adj	20,8	N Prep N Adj	37,9	N Prep N Prep N	27,6

Tabla 12: Comparación de soluciones entre el corpus paralelo y el diccionario Mosby.

Aunque las tendencias en las soluciones son claras, es importante tener en cuenta que los porcentajes de estos patrones pueden variar mucho ya que mientras una solución representa el total de las ocurrencias en el corpus paralelo, en el diccionario Mosby pueden sólo representar a un tercio de las soluciones. En principio, esta variación se debe al tamaño del corpus paralelo que no permite obtener más datos como ya se ha comentado. Sin embargo, si se comparan los patrones más frecuentes del corpus en inglés como es el caso de Adj Adj N, se verá que la diferencia no es enorme entre las soluciones del corpus paralelo y el diccionario Mosby (87,8 vs. 72,3).

Así, pues, las propuestas de otros autores sobre la solución de sintagmas nominales complejos del inglés al español (López y Minett 1997; Linder 2002 y Vivanco 1994) que se han presentado en §2.5 no dejan de ser propuestas muy

intuitivas a un problema que el traductor debe encarar a diario. Dichas propuestas no aportan más que confusión y ambigüedad en la interpretación y traducción de este tipo de sintagmas. Como se ha visto en este capítulo, existen regularidades no sólo en la frecuencia en el corpus paralelo en español sino en el diccionario Mosby. Así, pues, si un traductor quiere traducir algunas de las estructuras presentadas en este trabajo, que por su frecuencia son las que potencialmente tendrá que enfrentar en un trabajo diario, puede optar por observar las estructuras que aquí se proponen para solucionar este tipo de sintagmas. De igual modo, cuando tenga más de una posibilidad en español, los porcentajes encontrados le ayudarán a decidir cual estructura es la más prototípica para determinado patrón en inglés.

Si al consultar un diccionario encuentra igualmente más de una solución que esté determinada por una variación sintáctica, los datos de esta tesis le serán de suma ayuda a la hora de seleccionar uno de los equivalentes.

Asimismo, los datos de esta tesis pueden ser muy útiles al traductor, si después de agotar todas sus fuentes de consultar, no encuentra los equivalentes de un sintagma en inglés y necesita crear el término por completo o resolver el sintagma, pues a partir del análisis del patrón en inglés y de comparar las posibilidades que aquí se proponen puede construir un equivalente adecuado en español.

Del mismo modo, desde un punto de la enseñanza de la traducción, los datos de este capítulo son aún más relevantes si se tiene en cuenta que este tipo de sintagmas es una de las dificultades añadidas que debe enfrentar el estudiante de traducción y, que como se ha expuesto al comienzo de esta tesis, el profesor de traducción científico-técnica no tiene más elementos didácticos que su experiencia como traductor, si la tiene. Así el profesor de traducción del inglés al español, que trata aspectos microlingüísticos de los textos científico-técnicos, puede enseñar las estructuras más frecuentes que existen en inglés para este tipo de discurso, sus aspectos morfosintácticos y semánticos, las estructuras más prototípicas para cada patrón en inglés y como estas

estructuras están representadas en los diccionarios. Así el traductor o aprendiz de traductor tendrá más elementos de juicio para resolver este tipo de sintagmas que simplemente la intuición y las instrucciones presentadas por otros autores que no llevan a cabo constataciones de los datos.

8.5. Correlación entre el corpus paralelo y los corpus *Tècnic* del Iula y CREA de la RAE

Para contrastar las similitudes vistas entre las soluciones del corpus paralelo y las encontradas en el diccionario Mosby, se han comparado los patrones encontrados en las soluciones con los datos analizados en los corpus de *Tècnic* del Iula y el corpus CREA de la Real Academia Española.

Este contraste permitirá, por un lado, observar si la resolución de sintagmas del inglés al español presenta interferencias morfosintácticas, es decir, los patrones empleados son piezas poco regulares en español y, por ende, no se justifica su alta frecuencia en las soluciones presentadas en el corpus paralelo y, por otro lado, permitirá ver si los patrones de los sintagmas del diccionario Mosby siguen la intuición del hablante o igualmente están interferidos.

No son pocos los manuales de traducción y publicaciones que critican duramente las interferencias lingüísticas del traductor (Vázquez-Ayora 1977; García Yebra 1997; López y Minett 1997; Scarpa 2001; entre otros) que se ven reflejadas en calcos y anglicismos de todo tipo, en especial, léxicos y sintácticos. Así que las estructuras presentadas en el ítem anterior, y al igual que las soluciones encontradas en el diccionario, podrían tildarse de ser estructuras fruto de un calco sintáctico o estar interferidas lingüísticamente. Por eso, el uso de un corpus de contraste de lengua general, permitirá observar si dichas estructuras son posibles o no en español y si su frecuencia es igualmente alta.

Los datos de la tabla 13 listan los 20 patrones más frecuentes en los tres corpus analizados. Los datos de los patrones que se han contrastado del diccionario Mosby se presentan en la tabla 12.

Aunque solo se muestran los 20 primeros patrones de cada corpus, puede decirse que los patrones que sirven de soluciones en el corpus paralelo se encuentran todos en el corpus técnico aunque no con la misma frecuencia. De igual modo sucede con el corpus CREA de la RAE. En este caso, hay 4 patrones que no aparecen ya que no se incluyeron en el inventario de patrones para la consulta hecha a la RAE (N Prep N N, N Adj Adj Adj Adj, N Adj Adj Prep N y Adj N Adj) y, por tanto, no se puede decir que no sean estructuras propias del español. Los otros patrones se encuentran entre los 20 más frecuentes del CREA salvo los patrones N Adj Adj Adj y N Adj Prep N N que están situados de 33 y 23, respectivamente.

Tokens	Patrón corpus	Porc.	Tokens	Patrón Dic.	Porc.	Tokens	Patrón Crea	Porc.
3	N Prep N Adj	31,66	3	N Prep N Adj	15,17	3	N Adj Prep N	17,4
3	N Adj Prep N	16,13	3	N Adj Adj	10,64	3	N Prep N Adj	14,4
3	N Prep N Prep N	10,68	3	N Prep N Prep N	10,37	3	N Prep N Prep N	13,3
3	N Adj Adj	6,73	3	N Prep N N	7,76	3	N Prep N PP	10,4
3	N Adj PP	4,88	3	N Adj Prep N	7,67	3	Adj N Prep N	6,97
4	N Adj Prep N Adj	3,68	4	N Prep N Prep N Adj	3,76	3	N Adj Adj	5
4	N Prep N Adj Prep N	3,31	3	N N Adj	3,28	4	N Prep Adj N	3,9
3	Adj N Prep N	2,76	4	N Adj Prep N Adj	3,09	4	N Adj Prep N Adj	3,56
4	N Adj Prep N Prep N	2,67	3	N N N	2,67	3	N Prep N Adj Prep N	2,97
3	N Prep Adj N	1,84	4	N Adj Prep N Prep N	2,54	4	N Adj N	2,23
3	N PP Prep N	1,29	4	N Prep N Prep N Prep N	2,54	3	N Adj Prep N Prep N	2,22
4	N Prep N Prep N Prep N	1,11	4	N Prep N Adj Prep N	1,66	3	N Adj PP	1,95
4	N Prep N Prep N Adj	1,01	4	N Prep N Adj Adj	1,51	4	N N Prep N	1,95
3	N Prep N PP	0,93	3	N N Prep N	1,37	3	N Prep N Prep N Prep N	1,54
4	N Prep N Adj Adj	0,83	3	N Adj N	1,21	4	N N N	1,48
4	N Adj PP Prep N	0,74	3	N Prep Adj N	1,15	4	N N Adj	1,42
3	N N Adj	0,74	3	N Adv Adj	1	3	N Prep N Prep N Adj	1,23
5	N Adj Prep N Prep N Adj	0,64	4	N Prep N Adv Adj	0,74	4	N Adj Prep Adj N	1
3	N Adv Adj	0,65	4	N Adj Adj Adj	0,73	4	N Adv Adj	0,94
3	N Adj N	0,55	4	N Prep N Prep Adj N	0,57	3	N Adj PP Prep N	0,76

Tabla 13: Comparación de los primeros 20 patrones en español del corpus de referencia, el corpus paralelo y el corpus CREA.

Si se tienen en cuenta solo los patrones más frecuentes del corpus paralelo, puede observarse que igualmente son las cuatro estructuras más frecuentes del CREA. Así puede afirmarse que las estructuras empleadas por los traductores y en el diccionario Mosby son estructuras que no son fruto de interferencias lingüísticas y, por tanto, son perfectamente propias de español, y por extensión, del discurso especializado. En este sentido, no puede hablarse de calco de estructuras sintagmáticas del inglés al español sí, como sucede con el patrón Adj Adj N, un traductor emplea la estructura en español N Adj Adj que es el cuarto y sexto patrón más frecuente en el corpus de análisis y el CREA, respectivamente.

Es importante tener en consideración que no se han evaluado las traducciones ni los sintagmas nominales en cuanto a su corrección. Es posible que algunos de ellos estén errados o que haya mejores soluciones para un sintagma nominal. Sin embargo, puesto que las traducciones de la revista fueron realizadas por expertos o traductores profesionales y fueron sometidas a corrección de pruebas, asumimos que eran correctas. Lo que se quería verificar en este estudio es que ciertamente hay tendencias en la traducción de la premodificación compleja del inglés al español según lo hemos indicado. Este es un primer paso para buscar más regularidades, lo cual es muy relevante para los traductores y la formación de traductores, ya que no podemos confiar únicamente en la intuición del traductor.

8.6 Análisis de los patrones en inglés y los equivalentes en español

Como se ha podido observar los patrones que se han encontrado en español para uno en inglés no están interferidos según se constata de la comparación de los datos del corpus paralelo con los del corpus de análisis extraído del *Corpus Tècnic* del IULA, el diccionario Mosby y el CREA de la RAE.

A continuación, se presentan los patrones seleccionados en inglés y las soluciones de traducción al español más regulares del corpus paralelo.

Hay 13 patrones en inglés de 20 seleccionados como muestra que tienen soluciones regulares en español y se pueden dividir según su longitud, como se indica a continuación. La dependencia más frecuente en ambas lenguas se presenta dentro del patrón, de modo que sirve como forma de solución para ese patrón.

No todos los sintagmas de seis y siete tokens tienen una solución en español en patrones superficiales. Este hecho indica una alta variabilidad gramatical no sólo en inglés sino también en español.

Sólo hubo una solución regular en los sintagmas de 5 tokens. El patrón [Adj [Adj [Adj [Adj N]]]] tiene la misma solución en español (100%), [[[[N Adj] Adj] Adj] Adj]. Esto corresponde a una solución típica de derecha a izquierda, según se ilustra en 8.

8. symptomatic left ventricular systolic dysfunction [Adj [Adj [Adj [Adj N]]]]
disfunción sistólica ventricular izquierda sintomática [[[[N Adj] Adj] Adj] Adj]

Los otros patrones no presentaron ninguna regularidad al menos en términos de patrones superficiales.

Los patrones en inglés con una longitud de 4 tokens presentan varias soluciones. Los patrones Adj Adj Adj N y Adj Adj N N tienen las mismas soluciones en español N Adj Adj Adj (100%) y N Prep N Adj Adj (100%), respectivamente, como se ejemplifica en 9 y 10.

9. postoperative endoscopic retrograde cholangiography [Adj [Adj [Adj N]]]
colangiografía retrógrada endoscópica postoperatoria [[[N Adj] Adj] Adj]

10. central nervous system metastases [Adj [Adj N]] N]]

metástasis del sistema nervioso central	[N Prep [[N Adj] Adj]]

El patrón Adj N Adj N tiene dos soluciones N Adj Adj Adj con 4 ocurrencias (66,66%) y el patrón N Adj Prep Adj N con dos ocurrencias (33,33%), como puede observarse en los ejemplos 11, 12 y 13.

11.	large-scale clinical trial	[[Adj N] [Adj N]]
	ensayo clínico a gran escala	[[N Adj]] [[Prep Adj N]
12.	laparoscopic antegrade biliary stenting	[[Adj N] [Adj N]]
	endoprótesis biliar anterógrada laparoscopica	[[[N Adj] Adj] Adj]
13.	high baseline viral load	[[Adj N] [Adj N]]
	cargas virales básales elevadas	[[[N Adj] Adj] Adj]

El patrón Adj N N N tiene cuatro soluciones diferentes en español, como se indica en los ejemplos 14 a 17. Los patrones N Prep N Prep N Adj y N Adj Prep N Prep N tienen 4 ocurrencias cada uno (33,3%, respectivamente). Los patrones N Adj Prep N N y N Adj Adj Prep N tienen cada uno 2 ocurrencias (16,7%).

14.	global health status scale	[[Adj N] [N N]]
	escala del estado de salud global	[[N Prep [[N Prep N] Adj]]]
15.	systematic hypertension control programme	[[Adj N] [N N]]
	programa sistemático de control de la hipertensión	[[N Adj] [Prep N Prep N]]
16.	quantitative HCV RNA analysis	[Adj [N N] N]]
	análisis cuantitativos de ARN VHC	[[N Adj] [Prep N N]]
17.	stepwise Cox multivariate análisis	[Adj [N [N N]]]
	análisis multivariable gradual de Cox	[[[N Adj] Adj] Prep N]

El patrón N Adj N N con tres ocurrencias sólo tiene una solución (100%), como se ilustra en 18.

18.	Cox multiple regression analysis	[N [Adj N] N]]
	análisis de regresión múltiple de Cox	[[N [[Prep N Adj] Prep N]

El patrón N N N N con 7 ocurrencias tiene dos soluciones: N Adj Prep N N con 7 ocurrencias (71,42%) y N Prep N Adj con 2 ocurrencias (28,57%), como se muestra en 19 y 20.

19.	baseline HCV RNA concentration	[N [N N] N]]
	concentración basal de ARN VHC	[[N Adj] [Prep N N]]

20.	calcium channel blocker nimodipine	[[[N N] N] N]]
	antagonista del calcio nimodipino	[N Prep [N Adj]]

Con el patrón más productivo, Adj Adj N, 43 sintagmas (87,75%) de 49 ocurrencias se resolvieron con el mismo patrón en español (N Adj Adj), mientras que los otros seis sintagmas tienen tres soluciones diferentes. Esto sugiere una regularidad muy alta en este patrón. Este tipo de solución, en la cual el adjetivo relacional acompaña al núcleo en forma de compuesto sintagmático, ha sido analizado por otros autores (Crisma 1990; Zamparelli 1993)[69], quienes plantean que este tipo de estructura en lenguas romances con adjetivos relacionales se asemeja mucho al inglés. No obstante, las otras soluciones pueden ser muy útiles en los casos en los que en español se prefiere un sustantivo en vez de un adjetivo. Parece que en español, en el caso de un adjetivo derivativo de una parte del cuerpo, en español se prefiere el sustantivo de la parte del cuerpo, como se deduce del ejemplo 24[70].

21.	abnormal biological value	[Adj [Adj N]]

[69] Autores citados por Demonte (1999: 156).
[70] En algunos otros casos sucede precisamente lo contrario como lo indica el ejemplo 26.

	valor biológico patológico	[[N Adj] Adj]
22.	basic new fuchsin	[Adj [Adj N]]
	nueva fucsina básica	[[Adj N] Adj]
23.	centrilobular parenchymal damage	[Adj [Adj N]]
	lesión parenquimatosa centro lobular	[[[N Adj] Adj] Adj]
24.	cardiogenic pulmonary oedema	[Adj [Adj N]]
	edema de pulmón cardiogénico	[N [Prep N Adj]]

Además de la solución presentada en 21, existen otras tres soluciones, como se muestra en los ejemplos de 22 a 24: Adj N Adj, N Adj Adj Adj y N Prep N Adj, todas tres con dos ocurrencias, respectivamente.

La solución más común para el patrón Adj N N es N Prep N Adj con 27 (43,54%) de 62 sintagmas nominales, como en el ejemplo 25.

25.	anal canal dressing	[[Adj N] N]
	apósitos en el canal anal	[N [Prep N Adj]]

Las otras tres soluciones frecuentes para el patrón Adj N N que se ilustran de 26 a 29 son N Adj Adj con 12 ocurrencias (19,35%), N Adj Prep N con 9 ocurrencias (14,51%), N N Adj y N Adj, ambas con 5 ocurrencias (8,06%, respectivamente).

26.	systolic blood pressure	[Adj [N N]]
	presión arterial sistólica	[[N Adj] Adj]
27.	absolute neutrophil count	[Adj [N N]]
	recuento absoluto de neutrófilos	[[N Adj] Prep N]
28.	pathological Q wave	[Adj [N N]]
	onda Q patológica	[[N N] Adj]

29.	rheumatic heart disease	[Adj [N N]]
	cardiopatía reumática	[N Adj]

El patrón Adv Adj N con 3 ocurrencias tiene la misma solución en español, N Adv Adj (100%), como en 30.

30.	unexpectedly high response	[[Adv Adj] N]
	respuesta inesperadamente elevada	[N [Adv Adj]]

El patrón Adv PP N tiene la misma solución con una variante en el Adv Adj como en 31 y 32, ambos casos con 2 ocurrencias (50% en cada caso). Esta inversión es opcional en español debido a la nominalización de la oración.

31.	individually sealed envelopes	[[Adv PP] N]
	sobres cerrados individualmente	[N [Adj Adv]]

Este ejemplo puede interpretarse como *sobres que han sido separados de manera individual (o uno a uno)*.

32.	serologically proven influenza	[[Adv PP] N]
	gripe serológicamente demostrada	[N [Adv PP]]

En este caso, la interpretación del sintagma se origina en la oración *gripe que se ha demostrado mediante análisis serológicos*. Según lo afirma Gotti (2003: 70-71), este patrón puede originarse a partir de una voz pasiva (cuasi-pasiva en español) modificada por un adverbio, el cual va unido por un guión (no en todos los casos) al participio de pasado (o adjetivo deverbal en español) del verbo y colocado antes del sustantivo. En un sintagma más extenso, esto puede causar ambigüedades, lo que a su vez, puede causar problemas de lectura o traducción si las relaciones sintáctico-semánticas no se identifican adecuadamente.

El patrón N Adj N tiene principalmente la misma solución N Adj Adj con 6 ocurrencias de 12 (50%), como en el caso del ejemplo 32, pero también se encontraron los siguientes patrones: N Adj Prep N, N Prep N Adj y N Prep N Prep N, cada uno con 2 ocurrencias (16,66% en cada caso), como se ilustra en los ejemplos 33 a 35.

33.	brain natriuretic peptide	[N [Adj N]]
	péptido natriurético cerebral	[[N Adj] Adj]
34.	peak expiratory flow	[N [Adj N]]
	pico de flujo espiratorio	[N Prep [N Adj]
35.	chest radiographic findings	[N [Adj N]]
	hallazgos de la radiografía de tórax	[N Prep [N Prep N]]

El patrón N N N con 29 ocurrencias tiene principalmente estas dos soluciones: N Prep N Adj con 11 ocurrencias (37,93%) y N Prep N Prep N con 8 ocurrencias (27,58%). No obstante, hubo otras soluciones tales como N Adj Prep N y N Prep N N, ambas con 4 ocurrencias (13,79% en cada caso). Los ejemplos de este patrón se presentan en 36 y 37.

36.	aspartate aminotranferase concentration	[[N N] N]
	concentraciones de aspartato amino-transferasa	[[N Prep N] Adj]
37.	agarose gel electrophoresis	[[N N] N]
	elecfroforesis en gel de agarosa	[[N Prep [N Prep N]

El patrón PP Adj N con 6 ocurrencias tiene la misma solución N Adj PP, como en el ejemplo 38.

38.	isolated systolic hypertension	[PP [Adj N]]
	hipertensión sistólica aislada	[[N Adj] Adj]

También se pueden observar soluciones con respecto a las soluciones más frecuentes en español para un patrón específico en inglés. En este caso, la longitud del patrón en inglés no es relevante. Por ejemplo, el patrón en español N Adj Prep N es la solución de 10 patrones en inglés de longitud diferente (3, 4 y 5), como se enseña en los ejemplos 39 a 44.

39.	mean white blood cell counts	[Adj [[Adj [N N]] N]]
	recuentos medios de los leucocitos	[[N Adj] Prep N]
40.	baseline blood glucose concentrations	[N [N [N N]]]
	concentraciones basales de glucemia	[[N Adj] Prep N]
41.	absolute neutrophil count	[Adj [N N]]
	recuento absoluto de neutrófilos	[[N Adj] Prep N]
42.	individual dietary components	[Adj [Adj N]]
	componentes individuales de la alimentación	[[N Adj] Prep N]
43.	baseline HCV load	[N [N N]]
	carga basal de VHC	[N Adj] Prep N]
44.	allergen-specific immunotherapy	[[N Adj] N]
	inmunoterapia específica de alérgeno	[N Adj] Prep N]

Aunque no es el propósito de este capítulo el de analizar cada una de las soluciones del español con referencia de diferentes patrones en inglés, puede decirse que 15 patrones en español presentan el mismo comportamiento descrito antes para 19 patrones en inglés y 205 sintagmas involucrados.

8.7. Recapitulación

En este capítulo, se han presentado los resultados del análisis del corpus paralelo y el contraste de estos con el corpus de referencia *Tècnic* del IULA, CREA y el diccionario Mosby.

1. En cuanto a la longitud de los sintagmas, el corpus paralelo presenta las mismas tendencias que los otros corpus. Los patrones de 3 tokens predominan ampliamente sobre las de más longitud con un 64,3% de todas las ocurrencias.

2. En cuanto a la categoría gramatical predominante en la premodificación, se dan las mismas tendencias en el uso del sustantivo como premodificador por excelencia con un 32,56%, seguido por el adjetivo con un 32,23%. Las otras categorías léxicas tienen poca presencia en la premodificación. Sin embargo, dicho predominio es superior en los otros corpus.

3. En cuanto a los patrones más frecuentes, el corpus paralelo presenta las mismas tendencias de los otros corpus. Los patrones de Adj N N, Adj Adj N, N N N, entre otros, son igualmente muy frecuentes en los otros corpus.

4. De acuerdo con la dependencia sintáctica, la solución más común para los sintagmas de 3 tokens para la dependencia [C [B A]] son [[A B] C] con 86 ocurrencias y [A [B C]] con 27 ocurrencias y para la dependencia [[C B] A] es [A [B C]] con 39 ocurrencias. Para los sintagmas de 4 tokens existen 6 dependencias en inglés y 6 dependencias en español. La dependencia en inglés que presenta más regularidad es [[D C] [B A]] con la dependencia en español [[A B] [C D]] con 4 ocurrencias y para la dependencia en inglés [D [C [B A]]] con [[[A B] C] D] y [A [B C D]] ambas con 4 ocurrencias, respectivamente. Para

la dependencia [D [C B] A]] en inglés, la dependencia más regular en español es [[A B] [C D]] con 9 ocurrencias.

5. De acuerdo con el tipo de patrón, los patrones tienden a tener una sola solución en español en muchos casos o predomina al menos una de ellas. Así, los patrones Adj N N [N Prep N Adj], Adj Adj N [N Adj Adj], N N N [N Prep N Adj], N Adj N [N Adj Adj] y Adj N N N [N Prep N Prep N Adj] tienen las mismas soluciones o son las más frecuentes y representan el 62,4% de todas las ocurrencias.

6. Finalmente, las soluciones presentadas en cada patrón y, hasta cierto punto, su orden y frecuencia son las mismas que se encontraron en el diccionario Mosby. Igualmente, son los patrones más frecuentes en los corpus monolingües, excepto los casos mencionados. Así, puede afirmarse que existen regularidades en el comportamiento de las soluciones de este tipo de sintagma del inglés al español y que no existen diferencias importantes en el uso de estos sintagmas con respecto a lo que produce un experto en cada lengua, como se ha observado en los corpus monolingües.

9. Conclusiones: resultados y líneas de trabajo futuro

9.1 Síntesis de los resultados

La finalidad de esta tesis era, por un lado, demostrar tanto en el plano teórico como aplicado la existencia de los sintagmas nominales extensos especializados (SNEE) como característica relevante de las lenguas inglesa y española y, por el otro, proponer, en el plano aplicado, recomendaciones para el tratamiento desde el punto vista formal y semántico de estos sintagmas del inglés y sus correspondientes en español, para diferentes colectivos profesionales, en especial, para los traductores y terminólogos.

Para poder cumplir el primer objetivo y los objetivos específicos 1, 2, 3, 4, 6 y 7 se ha revisado la literatura en diferentes disciplinas del lenguaje para mostrar, en primer lugar, la existencia de muchos prejuicios frente a este tipo de estructuras y, en segundo lugar, presentar los autores que identifican o defienden los sintagmas nominales extensos como un rasgo del discurso especializado, en los cuales, se pretende compactar más información en poco espacio para hacer más eficiente y precisa la comunicación entre los expertos. En tercer lugar, hemos visto como un grupo de autores analiza los sintagmas nominales extensos desde una óptica didáctica tanto para la enseñanza de los lenguajes especializados como para la traducción. Sin embargo, sus propuestas para interpretar o traducir estos sintagmas no son suficientes ya que no tienen un planteamiento empírico que busque las tendencias que permitan dar soluciones eficientes desde un punto de vista didáctico y profesional. Igualmente, algunos autores han mostrado que este fenómeno es una característica de la lengua que se presenta con más frecuencia en el discurso especializado que en el discurso general.

Empíricamente, se ha creado un conjunto de patrones, se ha hecho una extracción en los dos corpus de referencia del IULA, y se han analizado cuantitativa y cualitativamente para caracterizar formal y semánticamente los

sintagmas nominales extensos especializados en ambas lenguas. Los resultados y conclusiones más importantes se describen a continuación.

1. Categoría léxica predominante en la premodificación y modificación: los resultados muestran que es el sustantivo, y no el adjetivo, la categoría por excelencia en ambas lenguas siendo más equilibrado el porcentaje en español debido a la estructura sintagmática de la lengua. Cabe destacar la presencia de los participios y los adverbios en ambas lenguas aunque en menor medida en español. En cambio, las categorías cerradas en español como las preposiciones representan un tercio de toda la modificación del sintagma. Esto se debe a que la expansión de sintagmas por posmodificación se hace mediante sustantivos que se encuentran dentro sintagmas preposicionales.

Por tanto, el uso predominante del sustantivo en la premodificación y modificación refuerza el carácter nominalizador del discurso científico-técnico. El sustantivo, como categoría que se refiere a entidades, sustancias, individuos, lugares y objetos más o menos concretos representa mejor las características del discurso científico-técnico y, por eso, su alta aparición en este tipo de discurso. Este aspecto favorece la "objetivización" del discurso debido al carácter estable, fijado y atemporal que proporciona la nominalización. Además, la necesidad de crear, nombrar o describir nuevos objetos, procesos y eventos en ciencia hace del sustantivo la categoría por excelencia no sólo como núcleo del sintagma sino como modificador en función adjetival o a través de sintagmas preposicionales, a pesar de que en lengua general es el adjetivo el modificador por excelencia.

En resumen, el sustantivo es la categoría léxica preferida en la premodificación y en la posmodificación en ambas lenguas y eso también se ve reflejado en la cantidad de patrones que no tienen adjetivos 42,46% contra los patrones que no tiene sustantivos (19,43%). Así, existe más del doble de patrones que no tienen adjetivos que aquellos que no tienen sustantivos, lo que también demuestra la preferencia del discurso científico-técnico por las nominalizaciones.

2. Longitud de los sintagmas: los patrones de 3 tokens predominan ampliamente sobre las otras longitudes en ambas lenguas, seguidos de lejos por los patrones de 4 tokens y, por último, los patrones de 5 y 6 tokens. Esta tendencia muestra que a mayor longitud del sintagma, menor frecuencia en un corpus. Aunque en esta tesis no se han trabajado sintagmas de 1 y 2 tokens, las tendencias muestran que su frecuencia debe ser inversamente proporcional a su longitud. Estos datos corroboran las afirmaciones de Quirk *et al* (1984: 1337-1338) en cuanto a la extensión de los sintagmas. Las consecuencias que se derivan de este hecho se resumen básicamente en las posibilidades que tiene un traductor de encontrarse este tipo de unidades y, en la confección de diccionarios especializados para determinar la cantidad de unidades que se deben incluir según la longitud.

3. Patrones más frecuentes del corpus: los patrones más frecuentes en inglés son N N N (30,05%), Adj N N (24,08%), Adj Adj N (10,71%) y el patrón N Adj N (5,88%). Estos cuatro patrones representan el 70,72% de todas las ocurrencias del corpus. De los patrones de 4 tokens, se pueden destacar los patrones Adj N N N y N N N N.

Además, se han clasificado los patrones por su extensión. Los patrones más frecuentes de 5 tokens son: Adj Adj N N N, Adj N N N N, N N N N N, Adj Adj Adj N N, Adv PP N N N; los patrones más frecuentes de 4 tokens Adj N N N, N N N N, Adj Adj N N, N Adj N N, Adj N Adj N, PP N N N, PP Adj N N; y los patrones más frecuentes de 3 tokens son: N N N, Adj N N, Adj Adj N, N Adj N, PP N N, PP Adj N, Adv Adj N, N PP N, Adv PP N y Adj PP N.

Si se comparan estos resultados a la luz de los resultados obtenidos por los autores que se han presentado en §2, puede verse que el patrón más frecuente N N N no aparece dentro de los patrones de co-ocurrencia

presentados por Biber *et al* (1999) y Montero (1995)[71] no lo presenta entre los más frecuentes de su corpus. Sin embargo, en el análisis contrastivo inglés-español que hace Montero es el menos frecuente de los analizados. En el caso del patrón Adj N N está descrito como uno de los patrones de co-ocurrencia más frecuentes en Biber *et al* y no aparece en Montero. El patrón Adj Adj N es el más frecuente en el corpus lexicográfico de Montero y, en Biber *et al* aparece sólo la forma de co-ocurrencia Adj Adj-color N que en nuestro caso no ha sido relevante. Finalmente, está el patrón N Adj N que no está descrito en Biber *et al* pero es el cuarto más frecuente de 3 tokens en Montero.

De los patrones más frecuentes de 3 tokens estudiados por Montero, el patrón Adv Adj N es el segundo más frecuente, pero en nuestro corpus sólo es el noveno más frecuente.

Todos los patrones presentados por *Biber et al* (1999) están dentro de los 13 más frecuentes de nuestro corpus (Adj N N, Adj Adj N, PP N N, Adv Adj N, Adv PP N, Adj PP N).

Biber *et al* (1999) no presenta patrones de 4 y 5 tokens. En cambio, Montero (1996) presenta 3 patrones de 4 tokens (Adj N Adj N, Adj Adj Adj N, N N Adj N) y 4 patrones de 5 tokens (Adv Adj Adj N, Adj Adj Adj N N, Adv PPi N N N, Adj N N Adj N) aunque sus frecuencias son demasiado bajas (1 patrón con 5 ocurrencias y el resto de 1 ocurrencia). De éstos, sólo los patrones Adj N Adj N y Adj Adj Adj N están dentro de los 20 más frecuentes de nuestro corpus. Los otros dos patrones, N N Adj N y Adv Adj Adj N, no están ni en los 10 más frecuentes de 4 tokens. El adjetivo que antecede al núcleo en estos dos últimos patrones está diferenciados en nuestro corpus con las categorías Adj y PP y, por tanto, tenemos los patrones N N PP N y Adv PP Adj N. Ninguno de los patrones

[71] En este sentido la estadística presenta en Montero (1995) no es clara ya que, por un lado, no hay un listado completo de patrones y, por otro lado, este patrón no aparece entre los más frecuentes aunque luego se afirme que representa el 20,21% de los patrones de 3 tokens (Montero 1995, 294).

de 5 tokens está entre los 20 más frecuentes de nuestro corpus de análisis. Sin embargo, los patrones Adj Adj Adj N N y Adv PP N N N son los dos menos frecuentes de los 5 encontrados en nuestro corpus y el patrón Adj N N Adj N no aparece en nuestro corpus.

En cuanto a los patrones más frecuentes en español, los resultados muestran los patrones N Prep N Adj (31,66%), N Adj Prep N (16,13%), N Prep N Prep N (10,68%), N Adj Adj (6,73%) y el patrón N Adj PP (4,88%). Estos cinco patrones representan el 70,08% de todas las ocurrencias del corpus y, por tanto, presentan menor variación sintáctica. Entre los patrones de 4 y 5 tokens cabe destacar los patrones N Adj Prep N Adj, N Prep N Adj Prep N y N Adj Prep N Prep N Adj.

Si se comparan nuestros resultados en español con los de otros autores, podemos ver que las estructuras más frecuentes de nuestro corpus están entre las más frecuentes en Cardero (2004) y Cartagena (1998), o son considerados patrones prototípicos por Vivaldi (2004).

De los 12 patrones estudiados por Cartagena (1998) solo 5 patrones están dentro de los 20 más frecuentes de nuestro corpus: N Prep N Adj Adj, N Prep N Adj Prep N, N Prep N Prep N Prep N, N Adj Prep N Prep N y N Adj Adj Prep N. Los otros 7 no se encuentran entre los más frecuentes: N Adj Adj Adj, Adj N Adj Adj, N Adj Prep N Adj, N Adj Adj Adj Prep N, N Adj Adj Prep Adj N y N Adj Prep N Adj Adj.

De los 4 patrones presentados por Vivaldi (2004) sólo el patrón N Adj Adj Prep N Prep N no aparece entre los 20 más frecuentes de nuestro corpus. Es importante resaltar que uno de los patrones más frecuentes en nuestros corpus, N Adj Adj, es presentado como uno de los más prototípicos por Vivaldi pero no por Cartagena (1998) ni por Cardero (2004).

De los 15 patrones estudiados por Cardero (2004) hay 10 patrones que aparecen entre los más frecuentes de nuestro corpus: N Prep N Prep N, N Prep

N Adj, N Prep Adj N, N Adv Adj, N Adj Prep N, N Adj N, N Prep N Prep N Prep N, N Prep N Prep N Adj y N Adj Prep N Prep N. Hay 5 patrones que no aparecen en nuestro corpus: N Adj Prep N Adj, N Adj Prep Adj N, N Adj Adj Adj, N Prep N Prep Adj N Prep N y N Prep Adj Conj Adj.

En resumen, se ha logrado compilar y clasificar cuantitativamente un número considerable de patrones que no han sido hasta ahora estudiados por otros autores en ambas lenguas, y se ha logrado contrastar la presencia de estos patrones en otros corpus para observar su frecuencia y uso y, se ha constatado que no existe una diferencia cuantitativa importante entre nuestros datos y los de los otros corpus.

Así, los resultados de los corpus de análisis permiten hacer las siguientes aseveraciones.

Los patrones más frecuentes de los "lenguajes de especialidad" no se diferencian de los patrones más frecuentes de los diccionarios ni de los corpus monolingües. Incluso en los pocos casos de patrones de esta tesis, que no aparecen dentro de los 20 más frecuentes del corpus CREA de la RAE, son patrones que están presentes dentro de la consulta general y se ubican en el rango de los patrones de mediana frecuencia. Lo que no se puede probar aquí es si este tipo de patrones es más frecuente en los discursos especializados que en lenguaje general ya que la consulta del CREA se hizo sobre todo el corpus. Es posible que la frecuencia sea mayor ya que los ejemplos proporcionados por la RAE tienden a ser especializados.

Sin embargo, lo que se pretendía observar era la exclusividad o no de estos patrones en los discursos especializados y si las explicaciones lingüísticas no se ajustaban a las de la lengua en general. Como consecuencia, estos patrones y las descripciones que se han hecho en §4 y §5 se han hecho desde la lengua general y no desde una perspectiva de una gramática de los "lenguajes especializados".

Se ha probado con el corpus lexicográfico conformado por diccionarios de diferentes áreas del conocimiento que la extensión de los sintagmas nominales complejos especializados no varía entre las ciencias “duras” y “blandas”, salvo en el caso del diccionario Routledge como se explicó en §5.4.1. Es decir, existe una relación directa entre la extensión del sintagma y la frecuencia de aparición en el diccionario. Tampoco existe diferencia alguna entre los patrones más extensos de las diferentes áreas.

4. Aunque se han compilado sintagmas de hasta 8 tokens, no se ha podido describir, clasificar, ni predecir el comportamiento de sintagmas de más de 6 tokens en los tres corpus de esta tesis y, por tanto, los resultados son parciales en este sentido. Aún así, no existe estudio alguno en las diferentes disciplinas relacionadas con el lenguaje que haya descrito, clasificado y explicado el comportamiento de sintagmas nominales extensos especializados en inglés y español de hasta 6 tokens.

Así, se ha demostrado que la existencia de los SNEE es una característica de la lengua que puede presentarse con mayor frecuencia en el discurso especializado, además, pueden describirse, clasificarse, explicarse y predecirse desde la gramática de una lengua como todos los fenómenos lingüísticos de los discursos de los ámbitos de especialidad, como lo plantea la teoría comunicativa de la terminología —TCT (Cabré 1999).

4. Relaciones de dependencia: la relación de dependencia [C [[B A]] es la más frecuente en todo el corpus en inglés con más del 60% de todas las ocurrencias del corpus de análisis, seguida de la relación de dependencia [[C B] A] con un 24,14% de todas las ocurrencias. Por último, la dependencia [[D C] [B A]] representa el 5,17% de todas las ocurrencias para patrones de 4 tokens. Los patrones que presentan una única relación de dependencias son: Adj Adj N, N Adj N, N PP N, PP Adj N, PP N N, Adv Adj N, Adj N N N y N N N N. De estos, los patrones Adj Adj N, N Adj N, N PP N, PP Adj N, PP N N tienen la misma relación de dependencia sintáctica [C [[B A]]. El único patrón que tiene la dependencia sintáctica [[C B] A] es Adv Adj N. En los dos patrones de 4 tokens,

Adj N N N y N N N N, la dependencia que predomina es [[D C] [B A]]. Los patrones que tienen dos relaciones de dependencia sintáctica son: Adj N N y N N N. En el patrón Adj N N, la relación de dependencia [C [[B A]] representa al 64,78% de las ocurrencias y [[C B] A] al 35,21%. En el caso del patrón N N N, la dependencia [[C B] A] representa al 87,93% y la dependencia [C [[B A]] al 12,07%.

En español, la relación de dependencia [A [B C]] es la más frecuente en todo el corpus con más del 50,5% de todas las ocurrencias del corpus de análisis, seguida de la relación de dependencia [[A B] C] con un 45,5%. Por último, la dependencia [[A B] [C D]] representa el 3% de todas las ocurrencias para patrones de 4 tokens. Los patrones que presentan una única relación de dependencias son: N Prep Adj N, N Adj Prep N, N Adj PP, N Adj Adj y Adj N Prep N y N Adj Prep N Adj. De estos, N Adj Prep N, N Adj PP, N Adj Adj y Adj N Prep N tienen la misma forma de dependencia sintáctica [[A B] C]. El único patrón que tiene la dependencia sintáctica [A [[B C]] es N Prep Adj N. El patrón de 4 tokens, N Adj Prep N Adj tiene la dependencia [[A B] [C D]]. Los patrones que tienen dos relaciones de dependencia sintáctica son: N Prep N Prep N y N Prep N Adj. En el patrón N Prep N Prep N, la relación de dependencia [A [[B C]] representa 70,37% de las ocurrencias y [[A B] C] al 29,62%. En el caso del patrón N Prep N Adj, la dependencia [A [[B C]] representa el 89,53% ocurrencias y [[A B] C] al 9,30%.

Las consecuencias que se derivan de estos resultados se pueden resumir básicamente en las posibilidades que tienen los traductores, estudiantes de inglés o español para propósitos específicos de interpretar estos sintagmas para su comprensión y traducción. Estas relaciones de dependencia muestran que la interpretación de sintagmas y sus posibles soluciones en otras lenguas no es lineal como lo han afirmado los autores presentados en §2.5. Igualmente, para la extracción de términos, las dependencias más comunes para determinado patrón permitirán dar un peso más específico a estos dentro de la extracción. Por ejemplo, los patrones de 4 ocurrencias presentan siempre la misma estructura de dependencia en forma binaria, lo que permite extraerlos de modo

confiable aunque no sean muy frecuentes. En el caso de patrones muy frecuentes como Adj Adj N en inglés con la misma dependencia sumarían un peso específico mayor que el anterior.

5. Clases semánticas: las clases semánticas más frecuentes de WordNet en el núcleo de los sintagmas del corpus de análisis son *noun.body* (18,53%), *noun.substance* (15,95%), *noun.act* (10,77%), *noun.group* (7,33%) y *noun.process* (7,33%). Estas cinco clases semánticas representan el 59,91% de todos los núcleos de la muestra. En UMLS, las clases semánticas más frecuentes son *Gene or Genome* (9,48%), *Biologically Active Substance* (9,05%), *Functional Concept* (6,03%), *Cell* (5,6%) y *Quantitative Concept* (4,74%). Estas cinco clases representan el 34,9% de todos los núcleos.

Las clases semánticas más frecuentes en la premodificación en WordNet son *noun.substance* (21,59%), *noun.body* (15,72%), *noun.animal* (11,53%), *noun.attribute* (9,01%) y, *noun.state* (4,4%). En UMLS las clases más frecuentes son *Gene or Genome* (9,48%), *Biologically Active Substance* (9,05%), *Functional Concept* (6,03%), *Cell* (5,6%) y *Quantitative Concept* (4,74%).

Obsérvese que WordNet 2.1 tiene más capacidad de generalización pero UMLS presenta más granularidad ya que las clases de los núcleos y la premodificación están más distribuidas entre las diferentes clases.

En español, las clases semánticas más frecuentes de EuroWordNet en el núcleo son *state* (21%), *act* (12%), *body* (11%), *cell* (7,5%) y *attribute* (7%). Estas cinco clases semánticas representan el 58,5% de todos los núcleos de la muestra.

Las clases semánticas más frecuentes en la modificación en EuroWordNet son *body* (24,88%), *adj.all* (17%), *substance* (14,04%), *state* (11,33%) y *cell* (10,1%). Estas cinco clases semánticas representan el 77,35% de toda la modificación de la muestra. Un aspecto importante en la modificación es que el promedio de adjetivos en cada posición tiende a aumentar de izquierda a

derecha, es decir, a medida que el modificador se aleja del núcleo existen más probabilidades de ser adjetivo.

6. Patrones semánticos más frecuentes: los patrones semánticos más frecuentes en inglés en WordNet son *animal notWN body* (2,59%), *animal substance body* (2,59%), *substance substance process* (2,16%), *body body substance* (1,72%) y *substance substance substance* (1,72%). Estos patrones semánticos obtenidos a partir de WordNet representan el 10,78% de la muestra de análisis. Los patrones más frecuentes en UMLS son CHEM CHEM CONC (3,45%), LIVB CHEM GENE (2,59%), CHEM CHEM CHEM (2,16%), CHEM CONC CONC (2,16%) y CONC ANAT ANAT (2,16%). Estos patrones semánticos obtenidos de UMLS representan el 12,52% de toda la muestra de análisis en inglés. Puede verse que en ambos programas no es posible obtener muchas generalizaciones en cuanto a los patrones ya que cada patrón semántico no abarca más del 3,5% de todas las ocurrencias en el mejor de los casos. Sin embargo, los patrones más frecuentes en ambos sistemas se correlacionan sintácticamente con los patrones superficiales más frecuentes tanto en el corpus de análisis en inglés como en el lexicográfico: N N N, Adj Adj N y Adj N N, N Adj N, PP N N y PP Adj N y en menor medida Adv Adj N. Puesto que los patrones semánticos tienen las clases semánticas más frecuentes, los patrones creados a partir de ellas y su asociación a los patrones superficiales más frecuentes muestra que son estas estructuras las más estables dentro de este estudio en todo sentido.

En español, los patrones más frecuentes en EuroWordNet son *state body all* (3,5%), *cell body body* (2%), *state body body* (2%), *act substance body* (1,5%) y *act substance substance* (1,5%). Al igual que en inglés, estos patrones semánticos tan solo representan el 10,5% de la muestra de análisis.

Puede verse que, al igual que en inglés, no es posible obtener muchas generalizaciones en cuanto a los patrones ya que cada patrón semántico no abarca a más del 3,5% de todas las ocurrencias en el mejor de los casos. Sin embargo, los patrones más frecuentes se correlacionan sintácticamente con dos

de los patrones superficiales más frecuentes tanto en el corpus de análisis en inglés como en el lexicográfico: N Prep N Adj y N Adj Adj/PP.

Los resultados reflejan lo "esperable" en cuanto a las clases semánticas puesto que el área temática de este estudio, el genoma, tiene involucradas estas clases antes presentadas. Por tanto, su aporte a este estudio es limitado. Sin embargo, el análisis realizado y la asociación que se ha hecho entre los patrones superficiales y los semánticos permiten saber que existe un uso adecuado entre los patrones y las clases semánticas de un área temática determinada, e. g., economía. Si se tiene en cuenta que algunas áreas del conocimiento están bien desarrolladas en determinadas ontologías, es posible trasladar los resultados de este estudio hacia campos de aplicación, como el etiquetaje de corpus, traducción automática, ontologías, extracción de terminología, lexicografía, etc.

Los aspectos semánticos dentro de un campo determinado pueden dar un peso adicional para extraer terminología. Por ejemplo, la suma de los factores siguientes con el patrón Adj Adj N permitirá extraer candidatos a términos con una mayor confiabilidad: patrón Adj Adj N, alta frecuencia, la misma dependencia sintáctica [C [B A]], asociación a patrones semánticos frecuentes CONC CONC CONC, CONC CONC DISO y CONC LIVB ANAT, adjetivos relacionales o paraterminológicos y un núcleo terminológico o paraterminológico.

En cuanto al corpus paralelo, se ha analizado cuantitativa y cualitativamente para caracterizar los sintagmas nominales extensos especializados en cuanto a las soluciones. Los resultados y conclusiones más importantes se describen a continuación.

1. Longitud de los sintagmas en el corpus paralelo: el corpus paralelo presenta las mismas tendencias que los otros corpus. Los patrones de 3 tokens predominan ampliamente sobre la demás longitud con un 64,3% de todas las ocurrencias.

2. Categoría gramatical predominante en la premodificación en el corpus paralelo: se dan las mismas tendencias en el uso del sustantivo como premodificador por excelencia con un 32,56%, seguido por el adjetivo con un 32,23%. Las otras categorías léxicas tienen poca presencia en la premodificación. Sin embargo, dicho predominio es superior en los otros corpus.

3. Patrones más frecuentes en el corpus paralelo: el corpus paralelo presenta las mismas tendencias de los otros corpus. Los patrones de Adj N N (21,77%), Adj Adj N (16,43%), N N N (11,89%), N Adj N (4,49%), Adj N N N (4,24%) son igualmente muy frecuentes en los otros corpus. Estos 5 patrones representan al 58,82% de ocurrencias del corpus paralelo. De acuerdo con la extensión de los sintagmas, los patrones más comunes de 3 tokens son: Adj N N, Adj Adj N, N N N, N Adj N, PP Adj N, PP N N, Adv Adj N, N PP N y Adv PP N; los de 4 tokens son Adj N N N, N N N N, Adj Adj N N, Adj Adj Adj N, N Adj N N, Adj N Adj N y N N Adj N; y los patrones de 5 tokens son Adj N N N N, N N N N N, Adj Adj N N N y Adj Adj Adj Adj N.

4. Relaciones de dependencia sintáctica en el corpus paralelo: la correspondencia de los patrones más común en los sintagmas de 3 tokens para la dependencia en inglés [C [B A]] son [[A B] C] del español con 86 ocurrencias y [A [B C]] del español con 27 ocurrencias y para la dependencia [[C B] A] es [A [B C]] del español con 39 ocurrencias. Para los sintagmas de 4 tokens existen 6 dependencias en inglés y 6 dependencias en español. La dependencia en inglés que presenta más regularidad es [[D C] [B A]] con la dependencia en español [[A B] [C D]] con 4 ocurrencias y para la dependencia en inglés [D [C [B A]]] con [[[A B] C] D] y [A [B C D]] ambas con 4 ocurrencias, respectivamente. Para la dependencia [D [C B] A]] en inglés, la dependencia más regular en español es [[A B] [C D]] con 9 ocurrencias.

5. Relación ente los patrones en inglés y los patrones equivalentes en español: los patrones en inglés tienden a tener un mismo patrón en español, donde –en muchos casos- predomina uno de ellos. Así, los patrones Adj N N [N

Prep N Adj], Adj Adj N [N Adj Adj], N N N [N Prep N Adj], N Adj N [N Adj Adj] y Adj N N N [N Prep N Prep N Adj] tienen los mismos patrones en español o son los más frecuentes y representan el 62,4% de todas las ocurrencias.

6. Relación de las soluciones del corpus paralelo y los corpus de contraste: los patrones en español presentados para cada patrón del inglés y, hasta cierto punto, su orden y frecuencia son los mismos que se encontraron en el diccionario Mosby. Igualmente, son los patrones más frecuentes en los corpus monolingües, excepto los casos mencionados. Así, puede afirmarse que existen regularidades en el comportamiento de las soluciones de este tipo de sintagma del inglés al español y que estas estructuras son igualmente frecuentes en la lengua general como se ha observado en los corpus monolingües.

Finalmente, los resultados del corpus paralelo permiten afirmar que:

1. hay regularidades en los patrones encontrados en inglés y español tanto en la extensión como en las estructuras que demuestran que los SNEE son estructuras de la lengua que pueden describirse dentro del marco gramatical de cada lengua de hasta 6 tokens.

2. existen regularidades en las soluciones de esos patrones en inglés al español de hasta 6 tokens. En 5 casos existe una única solución (62,4%) y en los otros 7 casos dos o más soluciones, pero con predominio de una de ellas.

3. no existen soluciones lineales como lo muestran las relaciones de dependencia. La pretendida modificación lineal propuesta por muchos autores de traducción (Linder, López y Minett, Vivanco, etc.) no existe siempre como tal. Las restricciones proporcionadas por los adjetivos de muchos patrones ayudan en la solución de muchos de ellos

4. las soluciones dadas por los traductores son las mismas que se han constatado en el diccionario Mosby del mismo ámbito temático que el corpus de

análisis, con lo cual no se puede acusar a los traductores de interferencia sintáctica.

5. los patrones empleados en las soluciones dadas por los traductores son estructuras frecuentes y propias del español como lo demuestra la comparación con el corpus CREA de la RAE.

6. el uso de corpus demuestra, una vez más, ser de gran utilidad para solucionar problemas lingüísticos de la traducción que no pueden dejarse simplemente a la intuición del hablante ni dar reglas de manera prescriptiva a un problema complejo, pues se corre el riesgo de no ser sistemático y cometer errores innecesarios. Por tanto, los estudios empíricos son de mucha utilidad para solucionar este tipo de problemas y retroalimentar los postulados teóricos de una disciplina.

7. el desarrollo de cualquier actividad científica debe sufrir el proceso normal de la ciencia: observar el objeto de estudio, describir su comportamiento, explicar y clasificar sus regularidades (y controlar las excepciones) y predecir el comportamiento del fenómeno en otras condiciones.

9.2 Validación o falsación de hipótesis

A continuación, se revisará la validación de las hipótesis a la luz de los resultados obtenidos, sin dejar a un lado las limitaciones que se han tenido y el alcance que éstos pueden tener.

1. Los sintagmas nominales extensos especializados no son un problema del discurso especializado, son un fenómeno de la lengua que presenta mayor frecuencia en el discurso especializado y que tiene unas características sintáctico-semánticas determinadas.

Está hipótesis se cumple plenamente ya que los resultados cuantitativos demuestran que este tipo de sintagmas está presente no sólo en los corpus de análisis del *Corpus Tècnic* del IULA sino también en los diferentes diccionarios de los corpus lexicográficos y la consulta hecha al corpus CREA de la RAE (empleados como corpus de contraste). La mayor presencia de estos sintagmas en el discurso especializado no es posible probarlo directamente ya que la consulta al corpus CREA de la RAE es general y no se ha separado por ámbitos ni niveles de especialidad. Sin embargo, los ejemplos proporcionados a través de la consulta al CREA pueden considerarse especializados y, por tanto, puede afirmarse que esta tendencia se presenta en una consulta más elaborada para los corpus monolingües generales. En inglés, no se ha hecho dicha consulta ya que los datos aportados por otros investigadores en este sentido confirman que su uso en el discurso especializado es mayor (Biber *et al* 1999).

El análisis ha demostrado que se han determinado las características de estos sintagmas desde la gramática de la lengua general, entre las cuales se pueden mencionar los patrones predominantes y su extensión explicados anteriormente, los sustantivos que pertenecen preferentemente a la clase de deverbales por sufijación y los adjetivos a la clase de denominales por sufijación, lo cual permite restringir también a determinados patrones. Además, se describió el conjunto de relaciones de dependencia asociadas a los patrones frecuentes del corpus y categorías semánticas que dependen del ámbito de estudio.

2. Los sintagmas nominales extensos especializados pueden describirse, clasificarse, explicarse y predecirse desde la gramática de una lengua como todos los fenómenos lingüísticos de los discursos de los ámbitos de especialidad.

Esta hipótesis se cumple en parte ya que se han podido describir, clasificar y explicar sólo los patrones más frecuentes de hasta 6 tokens si bien se han extraído sintagmas de hasta 8 tokens. Lo cierto es que no nos hemos puesto un límite máximo ya que teóricamente no existe, pero sí se ha establecido un límite mínimo como se ha afirmado en la introducción y la metodología. Así, se

han descrito y clasificado los patrones más frecuentes de los corpus, su frecuencia de acuerdo con la extensión, la presencia de la categoría léxica predominante en la modificación, las características y restricciones morfológicas de las diferentes categorías léxicas presentes en los sintagmas, las restricciones de las dependencias sintácticas en cada corpus de análisis en general y en cada patrón, las clases semánticas que predominan en los núcleos y en la modificación y los patrones semánticos obtenidos a partir de las clases y sus correlaciones con los patrones de superficie; todo dentro del marco formal y semántico de las gramáticas generales del inglés y el español.

Igualmente, se ha podido explicar y predecir el comportamiento de las descripciones antes descritas de los sintagmas nominales extensos especializados como fruto de una relación pragmática entre el emisor y el destinatario de un texto en el marco de los ámbitos de especialidad. La situación comunicativa de estos interlocutores y los objetivos que persiguen las ciencias permite explicar las características lingüísticas de este tipo de sintagmas y su función como un elemento que permite vehicular gran cantidad de información en poco espacio y, por eso se emplean determinadas estructuras de superficie y se privilegia el uso del sustantivo como categoría léxica y de determinado sufijos en las categorías léxicas estudiadas. Por eso, la extensión de los sintagmas y el fenómeno de nominalización presente son inherentes a la relación pragmática.

Para aumentar la cobertura de la validación de esta hipótesis sería necesario aumentar la cantidad de corpus, afinar los problemas de etiquetaje y la extracción que ya se han comentado en la metodología. De este modo, se obtendrían más sintagmas de mayor extensión y se podrían describir, clasificar y analizar formal y semánticamente. De todos modos, somos conscientes de que entre más extenso sea un sintagma más irregularidades se pueden esperar en el patrón de superficie lo que dificultaría su descripción desde este punto de vista. Aún así, pensamos que una descripción endógena por pares binarios del tipo Adv Adj, Adv PP, N N, Adj N, etc. como se ha hecho parcialmente para los sintagmas de 3 a 6 tokens puede ayudar en la descripción y clasificación de los sintagmas más extensos.

3. Existen regularidades en el comportamiento de las soluciones de traducción de este tipo de sintagma del inglés al español.

Esta hipótesis se cumple en parte ya que tiene las mismas limitaciones de la hipótesis anterior. Sin embargo, esta tesis ha probado que existe un comportamiento sistemático en los patrones equivalentes en español para determinado patrón en inglés como se observó en §8. A pesar de que no se encontraron regularidades importantes en patrones de más 5 tokens, las regularidades encontradas en los patrones de 3 a 5 tokens son suficientes para refutar las propuestas intuitivas hechas por los diferentes autores de traducción. Se han observado regularidades en el comportamiento de la extensión, restricciones de las dependencias sintácticas y el uso de un solo patrón en español para un patrón en inglés o, al menos el predominio de uno de ellos. Igualmente, las descripciones hechas parcialmente para los compuestos del tipo Adv Adj, Adv PP, N N, Adj N, etc. y sus correspondientes en español para los sintagmas de 3 a 5 tokens pueden ayudar a resolver patrones más extensos.

Por un lado, mediante la comparación del diccionario Mosby y el corpus paralelo y, posteriormente, con el corpus CREA de la RAE se ha podido observar si las soluciones encontradas en el corpus paralelo son idiosincrásicas o están interferidas por el inglés. La comparación de estas soluciones de los diferentes corpus muestra, por un lado, que este comportamiento es similar en el diccionario Mosby, es decir, que los mismos patrones en inglés tienen las mismas soluciones en español y tienden igualmente a ser la solución más frecuente del patrón en cuestión. Por otro lado, este comportamiento se ve corroborado en el corpus CREA ya que todos los patrones en español que se han empleado como soluciones a los patrones en inglés están presentes en la consulta. De hecho, son en su mayoría los más frecuentes de los analizados en el corpus CREA de los analizados, lo cual indica que son patrones propios del español y no están interferidos desde el punto de vista formal.

9.3 Aportes de la tesis

Los aportes de esta tesis pueden resumirse en las descripciones cuantitativas, formales y semánticas de los sintagmas nominales extensos especializados y sus aplicaciones en diversos campos del conocimiento relacionados con la lingüística: la gramática general, la traducción, la terminología, LSP y corpus.

9.3.1 Aportes sobre la descripción de los SNEE

Los resultados y las descripciones realizadas en esta tesis tienen implicaciones teórico-descriptivas en diferentes áreas del lenguaje.

9.3.1.1. Gramáticas de la lengua general

Parte de los resultados obtenidos en esta tesis pueden ser muy útiles en una gramática descriptiva.

Los resultados de la descripción formal de los sintagmas nominales de los capítulos §4 y §5, en especial, pero también la descripción semántica de los capítulos §6 y §7, demuestran que, al menos, las estructuras más frecuentes del corpus y las restricciones morfológicas y de dependencia sintáctica deben ser descritas en las gramáticas generales. Además, si se tiene en cuenta que la sintagmación es uno de los recursos sintácticos más frecuentes para formar nuevas unidades de significado, su descripción debería estar incluida en las gramáticas.

Así, la estadística sobre la frecuencia de los sintagmas de la lengua y los patrones más frecuentes de acuerdo con la extensión pueden ser útiles para explicar el uso de este tipo de sintagmas en los registros científico-técnicos, para dar cuenta de este tipo de estructuras en la lengua como fenómenos lingüísticos propios de la lengua.

Igualmente la clasificación y la descripción morfológica y de dependencias de los patrones más frecuentes que se ha hecho debe estar dentro de una gramática general. En primer lugar, la estadística de las categorías léxicas descritas en los núcleos y la modificación, la preferencia morfológica de los núcleos y la modificación de los sintagmas son relevantes para explicar las preferencias de ciertos tipos de discursos y explicar los fenómenos de nominalización presentes en la lengua. Igualmente las restricciones y asociaciones descritas parcialmente entre los compuestos Adv Adj, Adv PP, N N y Adj N permitirán explicar las relaciones internas de este tipo de sintagmas en la gramática y el papel que estas estructuras juegan en la lexicalización de sintagmas nominales extensos.

Si se observan las gramáticas del inglés, este tipo de estructuras sintagmáticas no está descrita adecuadamente, salvo las cuatro estructuras presentadas por Biber *et al* (1999), muchas de las más frecuentes no están descritas. Además debe tenerse en cuenta que su frecuencia es alta y que muchas tienden a representar objetos, procesos, entre otros que tienden a lexicalizarse.

En español no hay una sola gramática prescriptiva o descriptiva que explique este tipo de sintagmas en español. En principio, este tipo de estructuras presenta una frecuencia muy alta si se comparan con otros fenómenos menos frecuentes descritos en las gramáticas.

Por ejemplo, las estructuras N Prep N Adj, N Adj Prep N, N Adj Adj, N Adj Adv, N Adj PP, entre otras deberían tener una descripción en las gramáticas generales del español no solo como estructuras de la lengua general sino como estructuras que son frecuentes en los discursos especializados. En este sentido, algunos aspectos sobre las restricciones de los adjetivos en algunas de estas estructuras han sido parcialmente descritos por Demonte (1999) y Bosque (1999).

La falta de descripción de este tipo de estructuras sintagmáticas puede deberse a que en español los sintagmas nominales se expanden básicamente agregando complementos adjetivales o preposicionales. Sin embargo, no hay un inventario de la combinatoria de estos complementos en los sintagmas ni como se rigen las dependencias sintácticas dentro de ellos como se ha explicado en esta tesis.

9.3.1.2. Manuales de terminología

Si bien se reconoce la existencia de los sintagmas nominales extensos especializados en los diferentes manuales de terminología y de LSP, su descripción, como rasgo distintivo de los textos especializados y como problema terminológico, no está adecuadamente explicada. En primer lugar, ningún manual de terminología en español (Felber y Picht 1984, Cabré 1993, Arntz y Picht 1995, Fedor de Diego (1995) se explica la naturaleza, ni se clasifican ni explican las características lingüísticas de los sintagmas nominales especializados a pesar de que se reconoce que, en promedio, el 85% de los términos de un ámbito no son unidades simples. Si se tiene en cuenta este porcentaje, existe un gran vacío en este sentido ya que muchos de los términos compuestos o sintagmáticos son de más de 3 tokens.

Por tanto, los datos aportados por esta tesis no sólo en cuanto al corpus de análisis sino del corpus lexicográfico son un aporte en este sentido. Un manual de terminología debe ser capaz de describir cuantitativamente las estructuras más frecuentes de los diferentes corpus a nivel general, de acuerdo con la extensión y explicar su comportamiento lingüístico como se ha hecho en los capítulos §4 y §5. Igualmente debe dar cuenta de las categorías léxicas más frecuentes, qué características morfológicas presentan, cómo se asocian dentro de los sintagmas y que restricciones asignan a determinadas estructuras para ayudar en su lexicalización. En este sentido, esta tesis ha aportado los siguientes aspectos teórico-metodológicos.

1. se han analizado cuantitativamente las 20 estructuras más frecuentes en todos los corpus y se han comparado para observar las regularidades de los corpus y los diccionarios.

2. se han clasificado de acuerdo con su extensión y analizado cuantitativamente en cuanto a su frecuencia en los corpus.

3. se ha establecido que la extensión no está asociada al área temática sino al tamaño del corpus textual y lexicográfico ya que entre más entradas tenga un diccionario, menos entradas sintagmáticas tiene.

4. se ha destacado el uso de otras categorías no prototípicas, como los participios de pasado y de presente y los adverbios y, el papel que estas juegan dentro de los sintagmas.

5. se han descrito los sufijos más frecuentes en las diferentes categorías léxicas de los sintagmas y su naturaleza epistemológica dentro del discurso especializado.

6. se han establecido las relaciones de dependencia de todos los patrones del corpus de análisis de modo que permita a un terminólogo interpretar un término (Cabré 1993: 185).

7. se han establecido algunas equivalencias formales entre los patrones más frecuentes del inglés al español. Esto es un aporte interesante para la terminología bilingüe ya que le permitirá al terminólogo crear rápidamente sintagmas en español teniendo en cuenta los datos cuantitativos y cualitativos de esta tesis.

Así, se pueden suplir la falta de descripción y el tratamiento de unidades sintagmáticas en los manuales de terminología.

9.3.1.3 Aporte a la TCT

En esta tesis se ha partido de los principios teórico-metodológicos de la Teoría comunicativa de la terminología –TCT (Cabré 1999) para extraer, describir, clasificar y explicar los sintagmas nominales extensos especializados, ya que estas unidades son unidades del lenguaje. Pensamos que esta tesis hace varios aportes y refrenda otros aspectos teóricos y empíricos a la teoría.

1. se ha corroborado el principio de que todos los fenómenos léxicos y sintagmáticos del discurso especializado pueden explicarse desde la gramática de la lengua general.

2. Se ha corroborado el principio de *Condición de lenguaje natural* en cuanto a que se han descrito una serie de patrones sintagmáticos en ambas lenguas a partir de los datos recogidos de corpus y se han contrastado no sólo con un corpus lexicográfico sino con un corpus monolingüe para dar cuenta de que no existe diferencia en las estructuras descritas con las de la lengua general, como lo propone la TCT.

3. se han descrito las características morfológicas de las categorías léxicas de los patrones como en la lengua general, destacando la activación de los procesos de nominalización de los discursos especializados.

9.3.2 La aplicabilidad de la descripción de los SNEE

Las aplicaciones de esta tesis se pueden resumir básicamente en cuatro aspectos: la base de datos con los resultados para el inglés y el español, las recomendaciones para la enseñanza de la traducción y la enseñanza de la terminología y la extracción de la terminología.

9.3.2.1 La base de datos

En esta tesis se ha construido una base de datos en la que se incluyen todos los datos analizados en cuanto a los dos corpus de análisis con sus patrones superficiales, patrón de extracción con las restricciones, la dependencia sintáctica, las clases semánticas de los núcleos y la modificación, los patrones semánticos, su extensión, contexto completo, fuente, número de documentos en los que aparece el patrón, la frecuencia absoluta del patrón, su extensión, entre otros datos. En cuanto al corpus paralelo, se proporciona la extensión, el sintagma, el patrón respectivo y la dependencia sintáctica en inglés y, el sintagma, el patrón respectivo y la dependencia sintáctica en español. Los datos se pueden filtrar desde cualquier campo para obtener los resultados descritos y explicados en esta tesis. En resumen, se presentan 1.055 registros en inglés, 1.102 en español y 210 del corpus paralelo con todos los datos antes descritos.

Este recurso puede ser empleado principalmente para la enseñanza de traducción científico-técnica y la enseñanza de terminología para traductores u otros profesionales como se explica más adelante.

9.3.2.2 Recomendaciones para la enseñanza de la traducción

Si se tiene en cuenta que los sintagmas nominales extensos especializados son un problema frecuente en la traducción, se esperaría que el fenómeno estuviera descrito, explicado en los manuales de traducción y se sugieran estrategias para traducirlos del inglés al español. Al contrario, en ninguno de los manuales de traducción ni en artículos de traducción se describen y se sugieren estrategias. A partir de los resultados de esta tesis, un profesor de traducción puede proponer estrategias diferentes, diseñar guías didácticas para la enseñanza de la traducción de los SNEE y compilar material para enseñar a traducir sintagmas nominales extensos especializados.

A continuación, se sugieren algunas estrategias. Ante la presencia de un sintagma el alumno debe seguir la siguiente estrategia, que se deben complementar con las explicaciones de los principales patrones en §8.6. El aprendiz de traductor debe:

1. identificar los límites del sintagma: el aprendiz de traductor debe saber donde comienza y donde termina el sintagma. Debe identificar ante todo el núcleo y después los modificadores de derecha a izquierda hasta que encuentre un determinante, una preposición o un verbo, principalmente.

2. identificar el patrón superficial del sintagma.

3. comparar si el patrón identificado está descrito en nuestra base de datos.

4. independientemente de si aparece o no en nuestra base de datos, el alumno debe identificar las relaciones de dependencia del sintagma para poder interpretarlo de modo correcto en inglés.

5. comparar su análisis con las relaciones de dependencia de la base de datos para observar si se trata de la dependencia más común del patrón.

6. a continuación, analizar los patrones en español encontrados para el patrón en inglés y los relaciona con la dependencia sintáctica que ha analizado.

7. a partir de estos datos, interpretar los sintagmas en español y sugiere una traducción de acuerdo con el patrón más frecuente sin dejar de analizar las otras posibilidades que puedan existir.

8. Si fuera posible, guardar los datos en una base de datos terminológica asociada a una memoria, de modo que si vuelve a aparecer dicho sintagma pueda recuperar la traducción solucionada.

Igualmente, el profesor puede preparar sintagmas de la base de datos con las formas compuestas del tipo Adv Adj, Adv PP, N N y Adj N para observar lo siguiente:

1. identificar estas estructuras dentro del sintagma.

2. identificar las relaciones de dependencia que ocurren en los sintagmas.

3. interpretar estas relaciones de dependencia

4. interpretar las restricciones de tipo morfológico que ocasionan dentro del sintagma como se ha discutido en §4.5.2 y 5.5.2.

5. hacer las interpretaciones posibles de estos compuestos en español.

6. proponer traducciones al español de estos compuestos dentro del sintagma.

De este modo, el profesor de traducción cubre los sintagmas más extensos como se ha dicho antes.

Igualmente, el profesor puede hacer resúmenes de artículos de investigación de diferentes áreas para que el estudiante identifique este tipo de sintagmas, entre otros aspectos lingüísticos dentro de la fase de análisis del texto y luego los traduzca correctamente dentro de la fase de traducción.

9.3.2.3 Recomendaciones para la enseñanza de la terminología

No son pocos los autores que afirman que los términos polilexemáticos componen la mayor parte de los términos y de las nuevas denominaciones que se crean en un ámbito de conocimiento. Sin embargo, ningún manual de terminología o terminografía los describe, explica y sugiere estrategias para tratarlos adecuadamente. Sólo en algunos artículos de terminología se describen

algunas estructuras sintagmáticas como se ha comentado antes. Aún así, solo se han descrito unos cuantos patrones, pero no los más frecuentes. A partir de los resultados de esta tesis, un profesor de terminología puede crear diferentes estrategias y crear material para tratar terminográficamente este tipo de estructuras para la extracción y análisis de denominaciones. Por un lado, el profesor de terminología puede emplear los patrones y los diferentes análisis lingüísticos expuestos en §4. y 5; y, por otro lado, los datos de traducción de §8.6 pueden ser útiles para el tratamiento de terminología bilingüe. Así, el aprendiz de terminólogo debe:

1. identificar los límites del sintagma durante la fase de extracción de los candidatos a términos. El aprendiz de terminólogo debe saber donde comienza y donde termina el sintagma. Este "découpage" es un verdadero problema en las lenguas romances y para solucionarlo, se debe tener en cuenta las estructuras de la base de datos.

2. identificar el patrón superficial del sintagma teniendo en cuenta los 20 más frecuentes encontrados en los corpus de análisis y lexicográfico, para dar una puntuación a los más frecuentes al comparar si el patrón identificado está descrito en nuestra base de datos.

3. durante la identificación, tener en cuenta los tipos de núcleo más productivos morfológicamente para aumentar las posibilidades del candidato a término y asignar una puntuación si el núcleo tiene un sufijo muy productivo. Por ejemplo, aquellos sintagmas que tienen el sufijo *–ión* y sus alomorfos tienen más posibilidades de ser término que aquel sintagma con la misma estructura y similar modificación pero sin un núcleo nominalizado (*transmisión autosómica recesiva* vs. *forma autosómica recesiva*).

4. observar los aspectos morfológicos de la modificación y asignar una puntuación más alta a las estructuras más prototípicas que se han descrito. Por ejemplo, aquellas estructuras que tienen sufijos nominalizadores y adjetivos relacionales tendrán más posibilidades de ser términos. Por ejemplo, si un

sintagma tiene el patrón N Adj Adj y tiene un núcleo terminológico, es decir, que pertenece al campo en cuestión y, además los dos adjetivos son relacionales, entonces ese sintagma tiene más posibilidades de ser término (*gen supresor tumoral*). En cambio, si uno de los adjetivos es calificativo ese sintagma tiene menos posibilidades que el anterior (*insuficiencia renal crónica*).

5. independientemente de si aparece o no en nuestra base de datos, el alumno debe identificar las relaciones de dependencia del sintagma para poder interpretarlo de modo correcto como lo propone Cabré (1993: 185).

6. debe comparar su análisis con las relaciones de dependencia de la base de datos para observar si se trata de la dependencia más frecuente del patrón.

7. si hace terminología bilingüe, analizar los patrones en español encontrados para el patrón en inglés y relacionarlos con la dependencia sintáctica que ha analizado.

8. a partir de esta información, introducir los datos en una base de datos y asignarle un valor de acuerdo con los datos sugeridos en nuestra base de datos.

9.3.2.4 Recomendaciones para la extracción de la terminología

Al igual que las recomendaciones hechas para la enseñanza de la terminología, muchas de ellas también son útiles para la extracción de terminología. A continuación, se describen algunas estrategias para que un extractor identifique las unidades sintagmáticas que se han descrito en esta tesis.

1. tener en cuenta las estructuras más frecuentes que hemos descrito tanto para los corpus de análisis como para el corpus lexicográfico. En general, casi todas pertenecen a las reglas [(SN SA)$_{SN}$ (SP (SN SA))$_{SP}$ (SP (SN SA))$_{SP}$] y [(SN)$_{SN}$ (SP SN)$_{SP}$ (SP SN)$_{SP}$ SP SN)$_{SP}$ (SP (SN SA))$_{SP}$|(SP SN)$_{SP}$].

2. se debe agregar, si es posible, un analizador sintáctico para que especifique las relaciones de dependencia de los diferentes constituyentes y asignar una puntuación mayor a los sintagmas de acuerdo con los datos del 4.3.5 y 5. 3.5. La mayor parte de los patrones responden a las dependencias, [C [[B A]] en inglés y [A [B C]] en español.

3. programar para que dé mayor peso a aquellas estructuras que tienen como núcleo a un sustantivo del diccionario del sistema o que morfológicamente tengan sufijos nominalizadores.

4. igualmente, el sistema debe dar más peso a aquellas estructuras que tengan adjetivos relacionales y en especial a los que tengan los sufijos más frecuentes *-ico, -al (-ar), -nte, -ble, -eo, -ario, -ino, -ivo, –oso* y *-udo* y a las estructuras que tienen adverbios en *-mente* que indican área temática o formas adverbiales del latín como *in vivo*, *in vitro*, etc.

5. combinar la extracción con información semántica del área mediante el uso de diccionarios especializados u ontologías para que cada sintagma que tenga dicha información en su núcleo o parte de la modificación sea mejor puntuado.

Por tanto, si se combinan los patrones, la información sintáctica, morfológica y semántica, conjuntamente con la frecuencia y la dispersión del sintagma en el corpus podrá extraerse sintagmas terminológicos de más de 2 tokens sin demasiado ruido.

9.4 Limitaciones de la tesis y líneas de trabajo futuro

Esta tesis, como cualquier trabajo de investigación, tiene sus limitaciones. La limitación principal de esta tesis tiene que ver con la extracción insuficiente de patrones sintagmáticos de más de 6 tokens y su análisis

correspondiente, tal como se había planteado inicialmente. No quiere decir eso que no haya sintagmas más extensos de 6, 7 y 8 tokens en los corpus analizados. De hecho, en ensayos previos se han obtenido mejores resultados en este sentido debido a que se hizo una extracción manual. Esto se debe a que entre más extenso sea el sintagma más difícil es su extracción automática debido principalmente a problemas de etiquetaje y desambiguación como se ha mostrado. Así que sería necesario compilar grandes corpus para detectar sintagmas nominales más extensos.

De esta forma, se pueden obtener más ocurrencias en ambas lenguas para hacer estadísticas más confiables y comparar las regularidades en los casos de seis, siete o más tokens. Somos conscientes de que, al menos en español, no es fácil compilar tales corpus paralelos con textos en formato IMMRAD. Para hacer generalizaciones, es necesario aumentar astronómicamente el corpus y, aun así, es muy probable que no haya suficientes sintagmas nominales para poder generalizar. Puesto que es una limitación procedimental es preferible emplear otros procedimientos tal como analizar la estructura de los sintagmas en pares binarios, como se ha explicado antes.

La segunda limitación de esta tesis tiene que ver con el tipo de estadística realizada ya que se hizo una estadística descriptiva absoluta y no relativa frente al corpus ya que los datos de los diferentes corpus se trataron de igual manera. Sin embargo, la consulta hecha a la RAE, en la cual se tienen ambos tipos de estadística, conserva en general los primeros 20 patrones aunque su orden sí varía. Para ello, hubiera sido necesario obtener la dispersión de cada uno de los patrones pero no fue posible obtener esta medida por problemas técnicos.

En nuestros resultados puede observarse que hay patrones con frecuencias medias que aparecen en casi todos los documentos del corpus. Si se tuviera en cuenta la aparición en los diferentes documentos del corpus en los primeros 20 patrones existe una tendencia importante entre la frecuencia y la aparición de un patrón en los textos. Es decir, entre más frecuente sea un patrón, aparece más en los textos del corpus. Esta relación no se correlaciona

exactamente con cuatro patrones (Adj N N, Adv Adj N, N N N N y Adj Adj N N) en los cuales hay una ligera variación entre su frecuencia y un aparente porcentaje más bajo. Sin embargo, ese porcentaje no es dramáticamente bajo puesto que se encuentra dentro del rango de los 20 primeros.

La tercera limitación de esta investigación radica en la metodología de extracción ya que se ha optado por hacer una exploración piloto del corpus y obtener patrones de la bibliografía antes de la extracción final. Posteriormente, se ha visto que podía explorarse antes el corpus lexicográfico para posteriormente hacer la extracción final del corpus. Esto no hubiera cambiado necesariamente los datos de la tesis ya que los patrones más frecuentes aparecen en todos los corpus. Sin embargo, los datos pueden tener algunas variaciones en los patrones de mediana frecuencia ya que algunos patrones de los diccionarios no se exploraron en el corpus de análisis. Esto no se hizo pues los datos de los diccionarios se adquirieron posteriormente y se usaron como corpus de contraste.

Una última limitación es que los datos cuantitativos relacionados con las categorías léxicas se deben mirar a la luz de los estudios relacionados con modificación múltiple. Es posible que, si se hace un estudio que incluya patrones de 2 tokens, los datos puedan variar ya que el patrón N Adj es muy frecuente en la lengua.

Esta tesis propone varias líneas de trabajo, de las que se pueden mencionar las siguientes.

1. Para afinar las generalizaciones sobre los patrones de esta tesis, es necesario contrastar los resultados contra un corpus general como se ha hecho, pero en este caso es necesario excluir las áreas técnicas del corpus para observar si los datos de esta tesis son lo suficientemente discriminantes para el discurso especializado.

2. Aunque los resultados de los análisis no han proporcionado suficiente información para hacer generalizaciones sobre patrones semánticos, han sido útiles en cuanto a la información que se ha agregado a la muestra de modo que confirmará aspectos relacionados con la pertenencia de los sintagmas al área temática. Para propósitos terminológicos es necesario continuar investigando en otras áreas del conocimiento y contrastando el corpus de lengua general para establecer las tendencias que se han presentado en esta tesis. Este contraste permitirá saber si el uso de una ontología es lo suficientemente discriminante en los sintagmas para entrenar sistemas de extracción de terminología. Así, la combinación de patrones, frecuencia, dependencias, aspectos morfológicos y restricciones sintácticas e información semántica del área permitirá minimizar el ruido que pueden ocasionar algunos patrones.

Como se ha comentado en las limitaciones, sería interesante medir la dispersión de los patrones en un corpus, de manera que en las herramientas de explicitación de los corpus se puedan introducir parámetros de uso.

3. Un área de mucho interés es estudiar los patrones de colocaciones binarias dentro de la modificación ya que al actuar como un conjunto sintáctico y combinado con algunas características morfológicas (i. e. adjetivos relacionales) pueden ayudar en la detección de unidades y servir de material para el entrenamiento de traductores. En esta misma línea es necesario estudiar también aquellos patrones o sintagmas que tienen guiones o son irregulares para aportar más soluciones en la descripción y clasificación de SNEE.

4. Finalmente, otro aspecto interesante puede ser trabajar en la interpretación de las estructuras como lo propone Gotti (2003: 69-73), en el sentido de parafrasear oracionalmente estas estructuras nominales con el fin de hacer explícitas las relaciones semánticas entre los diferentes elementos. Este tipo de trabajo es de vital importancia para la enseñanza de la traducción y de lenguajes especializados.

Bibliografía

Abad Nebot; Ferraz M. A.; Torrego G. (1980). *Curso de lengua española. Orientación universitaria*. Madrid: Editorial Alhambra.

Abberton, Evelyn. (1977). "Nominal Group Premodification Structures". En: Bald, W-R.; Ilson, R. (Ed.). *Studies in English Usage: The Resources of a Present-Day English Corpus for Linguistic Analysis*. Frankfurt am Main: Peter Lang. 29-72.

Abril Martí, Isabel; Ortiz Urbano, Cocha. (1998). "Formación de intérpretes de conferencia en el ámbito biosanitario inglés/español- la experiencia de la Facultad de Traducción e interpretación de la Universidad de Granada". En: Fernández, Leandro; Ortega, Emilio (Coords.). *Traducción e interpretación en el ámbito biosanitario*. Granada: Comares.

Adelstein, Andreína. (1998). "Condiciones de la reductibilidad léxica de los sintagmas terminológicos". En: *Actas del VI Simposio Iberoamericano de Terminología RITerm*, La Habana. [CD-ROM].

Alarcos, Emilio. (1980). *Gramática Funcional del Español.* Madrid, Gredos.

Alcaraz Varó, Enrique. (2000). *El inglés profesional y académico*. Madrid: Alianza.

Alonso, Araceli; Cabré, Teresa; De Yzaguirre, Lluís; Tebé, Carles. (2002). "La utilización de corpus paralelos alineados en la docencia de la traducción y de los lenguajes de especialidad". En: Iglesias, L.; Doval, S. (Ed.). *Proceedings of the*

Second International Contrastive Linguistics Conference. Santiago de Compostela: Publicacións de la Universidade de Santiago de Compostela. 71-82.

Alvar Ezquerra, Manuel. (1993). *La formación de palabras en español*. Serie Cuadernos de lengua española. Madrid: Arcos.

Amaro de Melo, Bianca. (1998). *Las unidades terminológicas complejas en el áreas de telecomunicaciones*. Tesis de DEA no publicada. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir. Teresa Cabré y Enilde Faulstich.

Arntz, Reiner. (1982). "Methoden der fachsprachlichen Übersetzerausbildung im Sprachenpaar Spanisch und Deutsch". En: Rodríguez Richart, José; Thome, Gisela; Wilss, Wolfram (Ed.). *Fachsprachenforschung und -lehre. Schwerpunkt Spanisch*. Tubinga: Narr. 114-117.

Arntz, Reiner. (1993). "Terminological Equivalence and Translation". En: Sonneveld, Helmi; Loening, Kurt l. *Terminology: Applications in Interdisciplinary Communication*. Amsterdam: John Benjamins. 13-15.

Arntz, Reiner; Picht, Heribert. (1995). *Introducción a la terminología*. Madrid: Pirámide.

Assal, Allal; Delavigne, Valérie. (1993). "Découpage des unités terminologiques complexes: limite des critères linguistiques". En: *Actes de la 4ème journée Erla-Glat, "Langues de spécialité, outils et théories"*. Bretagne: ENST de Bretagne. 175-193.

Banks, David. (Ed.). (2001). *Le group nominal dans le texte spécialisé*. Paris: L'Harmattan.

Bark, Julia. (1980). *Let's Write English*. New York: Academic Press.

Barker, Ken. (1998). “A Trainable Bracketer for Noun Modifiers”. En: *AI*, 196-210.

Barker, Ken; Szpakowicz, Stan. (1998). “Semi-Automatic Recognition of Noun Modifier Relationships”. En: *Coling-ACL '98, Proceedings 36th Annual Meeting of the Association for Computational Linguistics*. Montreal: Coling. 96-102.

Bauer, Laurie. (1982). *English Word-Formation*. Cambridge: Cambridge University Press.

Bédard, Jean-Claude. (1986). *La traduction technique: principes et pratique*. Montreal: Linguatech.

Bennett, Paul. (1993). “A Multilingual Translation-Oriented Typology of Compound Nouns”. En: *T.A.L.*, 34 (2). 48-58.

Bevilacqua, Cleci. (2004). *Unidades fraseológicas especializadas eventivas: descripción y reglas de formación en el ámbito de la energía solar*. Tesis doctoral. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir. Teresa Cabré.

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, Edward. (1999). *Logman Grammar of Spoken and Written English*. London: Logman.

Blake, Gari; Bly, W. Robert. (1993). *The Elements of Technical Writing*. New York: MacMillan.

Bodenreider, Olivier; McCray, Alexa. (2003). “Exploring Semantic Groups through Visual Approaches”. En: *Journal of Biomedical Informatics*, 36. 414–432.

Bosque, Ignacio. (1999). “El sintagma adjetival: Modificadores y complementos del adjetivo. Adjetivo y participio”. En: Bosque, Ignacio; Demonte, Violeta. *Gramática descriptiva de la lengua española*, 1. Madrid: Espasa. 217-310.

Bosque, Ignacio; Demonte, Violeta. *Gramática descriptiva de la lengua española*, 1. Madrid: Espasa.

Boughedaoui, Mourand. (1995). “Les séquences collocationnelles et la dynamique de la composition adjetivale”. En: *Les Cahiers de l'Apliut*, 15 (1). 47-58.

Boughedaoui, Mourand. (1996). “Essai de categorisation sémantique des adjetives composés (1)”. En: *Les Cahiers de l'Apliut*, 16 (1).

Boughedaoui, Mourand. (1996). “Essai de categorisation sémantique des adjetives composés (2)”. En: *Les Cahiers de l'Apliut*, 16 (2). 37-54.

Boughedaoui, Mourand. (1997). “Contribution à l’amélioration de la compréhension et de la traduction des adjectifs composés en classe de langue de spécialité”. En: *Asp*, 15-18. 225.

Boughedaoui, Mourand. (1998). “Comparative Study of the Distribution of Adverb-Adjective Combinations with a Special Concern in English for Statistics”. En: *Les Cahiers de l'Apliut*, 17 (2). 37-54.

Boughedaoui, Mourand. (2001). “Contribution des associations syntagmatiques adjectivales à la complexification du groupe nominal dans le texte spécialisé”. En: Banks, David (Ed.). *Le group nominal dans le texte spécialisé*. Paris: L’Harmattan. 137-150.

Bourigault, Didier. (1993). “Analyse syntaxique locale pour le reperage de termes complexes”. En: *T.A.L.*, 34 (2). 105-117.

Brown, Peter; Lai, Jennifer; Mercer, Robert. (1991). "Aligning Sentences in Parallel Corpora". En: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley: University of California. 169-176.

Burgos, Diego. (2006). "Concept and Usage-Based Approach for Highly Specialized Technical Term Translation". En: Gotti, M.; Sarcevic, S. (Ed.). *Translation of Specialized Text*. Linguistic Insights series. Berna: Peter Lang.

Burgun, Anita; Bodenreider, Olivier. (2001). "Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System". En: Proceedings of NAACL'2001 Workshop. Association for Computational Linguistics. 77-82.

Burnett, Rebeca. (1992). *Technical Communication*. Californa: Wadsworth.

Cabré, María Teresa. (1993). *La terminología: teoría, metodología, aplicaciones*. Barcelona: Antártida. [trad. Tebé, Carles].

Cabré, María Teresa. (1999). "Elementos para una teoría de la terminología: hacia un paradigma alternativo". En: Cabré, María Teresa. *La Terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. 69-92.

Cabré, María Teresa. (1999). "Hacia una teoría comunicativa de la terminología: aspectos metodológicos". En: Cabré, María Teresa. *La terminología: Representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 129-150.

Cabré, María Teresa. (1999). "Una nueva teoría de la terminología: de la denominación a la comunicación". En: Cabré, María Teresa. *La Terminología: representación y comunicación. Elementos para una teoría de base*

comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. 109-128.

Cabré, María Teresa. (1999). *La terminología: representación y comunicación. Una teoría de base comunicativa y otros artículos.* Barcelona: Institut Universitari de Lingüística Aplicada- Universitat Pompeu Fabra (Series monografies, 3).

Cabré, María Teresa. (2002). "Terminología y lingüística: la teoría de las puertas". En: *Estudios de Lingüística Española (ELIES)*, 16. [http://elies.rediris.es/elies16/Cabre.html].

Cabré, María Teresa. (2003). "El lenguaje científico desde la terminología". En: Gutiérrez Rodilla, Bertha M. (Ed.). *Aproximaciones al lenguaje de la ciencia*. Burgos: Fundación Instituto Castellano y Leonés de la Lengua. 19-52.

Cabré, María Teresa; Estopà, Rosa. (2005). "Unidades de conocimiento especializado, caracterización y tipología". En: Cabré, M. Teresa.; Bach, Carme. *Coneixement, llenguatge i discurs especialitzat*. 69-94.

Café, Ligia. (1999). *La Description et l'analyse des unités terminologiques complexes en langue portugaise (variété brésilienne)*. Tesis doctoral no publicada. Université Laval. Dir. Auger, Pierre; Faultish, Enilde.

Calonge, Julio. (1995). *El lenguaje científico y técnico*. En: Seco, M.; Salvador, G. (Ed.). *La lengua española, hoy*. Madrid: Fundación Juan March. 175-186.

Cardero García, Ana María. (2004). *Lingüística y terminología*. México: Facultad de Estudios Superiores-Acatlán, Universidad Nacional Autónoma de México.

Cardero, Ana María. (2000). “En torno a la frecuencia de algunas estructuras sintácticas en terminología”. En: *Actas de VIII Simposio Internacional de la Red Iberoamericana de Terminología*. Lisboa: Colibrí.

Cartagena, Nelson. (1998). “Acerca de la variabilidad de los términos sintagmáticos en textos españoles especializados”. En: Wotjak, Gerd (Ed.). *Estudios de fraseología de español actual*. Madrid y Frankfurt: Iberoamericana. 281-296.

Casadei, Federica. (1994). “Il lessico nelle strategie di presentazione dell’informazione scientifica: il caso della fisica”. En: De Mauro, T. (Ed.). *Studi sul trattamento linguistico dell’informazione scientifica*. Roma: Bulzoni. 47-69.

Chambers, Chris. (1994). “Analysing and Generating English Compound Structures for Machine Translation”. En: Bouillon, P.; Estival, D. (Ed.). *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*. Geneva: ISSCO. 125–134.

Chang, Jason; Chen, Mathis. H. (1997). “An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques”. En: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco, Madrid: UNED.

Coates, Jennifer. (1977). “A Corpus Study of Modifiers in Sequence”. En: Bald, W-R.; Ilson, R. (Ed.). *Studies in English Usage: The Resources of a Present-Day English Corpus for Linguistic Analysis*. Frankfurt am Main: Peter Verlag. 9-27.

Collet, Tanja. (1997). “La réduction des unités terminologiques complexes de type syntagmatique”. En: *Meta*, 42 (1). 193-206.

Collet, Tanja. (2003). "A Two-Level Grammar of the Reduction Process of French Complex Terms in Discourse" En: *Terminology* 9 (3). 1-27.

Cortés, Jesús Andrés. (2004). *Análisis lingüístico de términos comparados en inglés técnico agrícola*. Tesis de doctorado no publicada, Universidad Complutense de Madrid. Dir. María Luisa Vega.

Coseriu. Eugenio. (1973). *Einführung in die strukturelle Betrachtung des Worschatzes*. Tubinga: Gunter Narr.

Daille, Béatrice; Dufour-Kowalski, S.; Morin, E. (2004). "French-English Multi-Word Terms Alignment Based on Lexical Content Análisis". En: *Proceedings Fourth International Conference on Language Ressources and Evaluation (LREC 2004)*, 3. 919-922.

David, Sophie. (1993). *Les unités nominales polylexicales: Éléments de description et reconnaissance automatique*. Tesis doctoral no publicada. Paris: Université Denis Diderot. Dir. F. Corblin.

De Mauro, Tullio (Ed.). (1994). *Studi sul trattamento linguistico dell'informazione scientifica*. Roma: Bulzoni.

Demonte, Violeta. (1999). "El adjetivo: clases y usos. La posición del adjetivo en el sintagma nominal". En: Bosque, Ignacio; Demonte, Violeta. *Gramática descriptiva de la lengua española*, 1. Madrid: Espasa. 129-216.

Dixon, Robert. (1977). "Where Have All the Adjectives Gone?" En: *Studies in Language*, 1. 19-80.

Downing, Pamela. (1977). "On the Creation and Use of English Compounds Nouns". En: *Language*, 53 (4). 810-842.

Drouin, Patrick. (1997). “Une méthodologie d'identification automatique des syntagmes terminologiques: L'apport de la description du non-terme”. En: *Meta,* 42 (1). 45-54.

Durieux, Christiene. (1988). *Fondament didactique de la traduction technique*. Paris: Didier Erudition.

Durieux, Christine. (1997). “La Recherche terminologique en traduction: pour une approache hypertextuelle”. En: *Meta,* 42 (4). 677-684.

Escandell Vidal, M. Victoria. (1995). *Los complementos del nombre*. Madrid: Arco/Libros.

Estopà, Rosa. (1999). *Extracció de terminologia: elements per a la construcció d'un SEACUSE*. Tesis Doctoral. Dir. Teresa Cabré. Institut Universitari de Lingüística Aplicada: Barcelona.

Estopà, Rosa. (2000). “Los adjetivos en las unidades terminológicas poliléxicas: un análisis morfosemántico”. En: *Organon*, 14, (28/29). 233-246.

Estopà, Rosa. (2001). “Les unités de signification spécialisées: élargissant l’objet du travail en terminologie” En: *Terminology* 7 (2), 217–237.

Estopà, Rosa; Lorente, Mercè; Folguerà, Rosa-Anna. (2002). “El rol de los adjetivos en los textos especializados”. En: *Actas del VIII Simposio Iberoamericano de Terminología*. [CD-ROM].

Estopà, Rosa; Valero, Tony. (2002). “Adquisición de conocimiento especializado y unidades de significación especializada en medicina”. En: *Panace@ - Boletín de Medicina y Traducción*, 3 (9-10), 72-82. [http://www.medtrad.org/panacea].

Fähndrich, Ursula. (2005). "Terminology Project Management". En: *Terminology* 11 (2). 225-261.

Faultisch, Enilde. (2003). "Formação de termos: do constructo e das regras às evidências empíricas". En: Faulstich, Enilde; Pereira de Abreu, Sabrina. (Ed.). *Lingüística Aplicada à Terminologia e à Lexicografia*. Porto Alegre: UFRGS. 11-31.

Fedor de Diego, Alicia. (1995). *Terminología: teoría y práctica*. Caracas: Equinoccio.

Felber, Helmut; Picht, Heribert. (1984). *Métodos de terminografía y principios de investigación terminológica*. Madrid: Instituto Miguel de Cervantes.

Feliu, Judit. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Tesis doctoral. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir. Teresa Cabré.

Fernández, Salvador. (1986). *Gramática Española*. Madrid: Arco/Libros.

Finin, Timothy W. (1980). *The Semantic Interpretation of Compound Nominals*. University of Illinois, Urbana-Champaign. University Microfilms International.

Finin, Timothy W. (1986). "Contraining the Interpretation of Nominal Compounds in a Limited Context". En: Grishman, Ralph; Kittredge, Richard (Ed.). *Analysing Language in Restricted Domains*. New Jersey: Lawrence Erlbaum Associates. 163-173.

Folguerà, Rosana. (2002). *Adjectius en el discurs espacialitzat: Una pirmera descripció dels adjectius en els textos del Genoma Humà*. Tesis de DEA sin publicar. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir. Mercè Lorente.

Gale, William; Church, Kenneth. (1991). “A Program for Aligning Sentences in Bilingual Corpora”. En: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Morriston, NJ.: University of California. 177-184.

Gallardo San Salvador, Natividad. (1997). *El orden de la descripción de las características y su importancia para la denominación y traducción de un término: Casos que se presentan en términos de nutrición*. Tesis doctoral.. Departamento de Filología Inglesa, Facultad de Filosofía y Letras, Universidad de Granada. Dir. Fernando Serrano V.

Gallegos Shibya, Alfonso. (2000). “Morfología y registro: Algunas relaciones entre tradiciones discursivas y morfología derivativa en español”. En: *Función*, 20-24. 142-215.

Gallegos Shibya, Alfonso. (2003). *Nominalización y registro técnico. Algunas relaciones entre morfopragmática, tradiciones discursivas y desarrollo de la lengua en español*. Tesis doctoral no publicada. Fakultät der Albert-Ludwigs-Universität Freiburg i Br. Dir. Elisabeth Cheauré.

Galve, Ignacio Guillén. (1998). “The Textual Interplay of Grammatical Metaphor on the Nominalizations Occurring in Written Medical English”. En: *Journal of Pragmatics*, 30 (3). 363-385.

García Yebra, Valentín. (1989b)[1997]. *Teoría y práctica de la traducción*, 2ª. Ed., II Vol. Madrid: Gredos.

Georges, Thomas. (1996). *Analytical Writing for Science and Technology*. [http://mywebpages.comcast.net/tgeorges/write/] [Consultado el 15 de julio de 2003].

Gili Gaya, Samuel. (1961). *Curso Superior de Sintaxis Española*, Barcelona, Biblograf.

Giraldo Ortiz, John Jairo. (2005) *Análisis y descripción de las siglas en el discurso especializado de genoma humano y de medio ambiente*. Tesis de DEA no publicada. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir.: M. Teresa Cabré.

Gotti, Maurizio. (1991). *I Linguaggi specialistici: caratteristiche linguistiche e criteri pragmatici*. Firenze: La Nuova Italia.

Gotti, Maurizio. (2003). *Specialized Discourse: Linguistic Features and Changing Conventions*. Linguistic Insights, 8. Berna: Peter Lang.

Guzmán, Blanca Mercedes. (2002). *Compuestos nominales del discurso científico escrito en inglés (microbiología médica): un estudio retórico-terminológico*. Tesis de maestría no publicada. Facultad de Humanidades y Educación, Universidad de los Andes. Dir. Dr. Françoise Salager Meyer.

Halliday, Michael M. K. (1998). "Things and Relations: Regrammaticising experience and technical knowledge". En: Martin, J.R. and Veel, Robert (Ed.). *Reading Science: Critical and Functional Perspective on Discourses of Science*. London: Roudledge.

Halliday, Michael M. K.; Hasan, Ruqaiya. (1976). *Cohesion in English*. London: Logman.

Halliday, Michael. M. K. (1991). *An Introduction to Functional Grammar*. London: Edward Arnold.

Hanns, Michael. (1990). *The Key to Technical Translation*. Vol I y II. Amsterdam/Philadelphia: John Benjamins.

Herzog, Robert (1971). “Gengenwartige Tendenzen in der terminologischen Wortbildung”. En: *Mitteilungsblatt für Dolmetscher und Übersetzer*, 17 (9-10). 3-6.

Herzog, Robert. (1978). “On the Relative Order of Adjectives”. En: Seiler, H. (Ed.). *Language Universals*. Tübingen: Gunten Narr. 165-184.

Hoffmann, Lothar. (1987) [1975]. *Kommunikationsmittel Fachsprache*: *Eine Einführung*. Tübingen Gunter Narr: Tübinga.

Hoffmann, Lothar. (1998). *Llenguatges d'especialitat: selecció de textes*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Hölz-Mantari, Justa. (1984). *Translatorishes Handeln*. Helsinki: Soumalainen Tiedeakatemia.

Horsella, María; Pérez, Fresia. (1991). “Nominal Compounds in Chemical English Literature: Towards an Approach to Text Typology” En: *English for Specific Purposes*, 10. 125-138.

Huddleston, Rodney; Pullum, Geoffrey. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hurford, Jim. (1998). “The interaction between numerals and nouns”. En: Plank, F. (Ed.). *Noun Phrase Structure in the Languages of Europe*. Berlín: Walter de Gruyter. 561-620.

Iturrioz Leza, José Luis. (2000). “Diversas Aproximaciones a la nominalización. De las Abstracciones a las macrooperaciones textuales”. En: *Función*, 20-24. 32-140.

Jastrab de Saint-Robert, Marie-Josée. (1987). "Les syntagmes nominaux complexes en anglais et en française". En: *Meta*, 32 (3). 260-266.

Kaul de Marlangeon, Silvia Beatriz. (2002). *Los adverbios en -mente del español de hoy y su función semántica de cuantificación*. Frankfurt am Main: Vervuert/Madrid: Iberoamericana.

Kirkman, John. (1992). *Good Style: Writing for Science and Technology*. London: Chapman & Hall.

Kocourek, Rostilav (1981). "Prerequisites for an Applicable Linguistic Theory of Terminology". En: Savard, Jean-Guy; Laforge, Lorne. *Actes du 5e Congrès de l'Association Internationale de Linguistique Appliquée*. Québec: Presses de l'Université Laval. 216-228.

Kocourek, Rostilav. (1991). *La langue française de la technique et de la science: vers une linguistique de la langue savante*, 2.º ed. Wiesbaden: Brandstetter.

Kornfeld, Laura; Resnik, Gabriela. (2002). "Sintagmas terminológicos con adjetivos pasivos". En: *Actas del VIII Simposio Iberoamericano de Terminología: La Terminología, entre la globalización y la localización*. Cartagena, Colombia. [http://www.riterm.net/actes/8simposio/indice02.htm].

Lacuesta, Ramón; Bustos, Eugenio. (1999). "La derivación nominal". En: Bosque, Ignacio y Demonte, Violeta. *Gramática descriptiva de la lengua española*, 3. Madrid: Espasa. 4505-4594.

Ladouceur, Jacques; Drouin, Patrick. (1997). "Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité". En: *Meta*, 42 (1). 207-218.

Lagoudaki, Elina. (2006). *Translation Memory Survey 2006*. Imperial College London: London. [http://www3.imperial.ac.uk/portal/pls/portallive/docs/1/7307707.pdf].

Le Masle, Karine. (2001). "Syntagme nominal fleuve dans le droit de l'environnement: la désignation des déchets". En: Banks, David (Ed.). *Le group nominal dans le texte spécialisé*. Paris: L'Harmattan. 65-72.

Lehrberger, John. (1982). "Automatic Translation and the Concept of Sublanguage". En: Kittredge, Richard; Lehrberger, John (Ed.). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: Walter de Gruyter. 81-106.

Leonard, Rosemary. (1984). *The Interpretation of English Noun Sequences on the Computer*. The Netherlands: Elsevier.

Levi, Judith N. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

L'Homme, Marie-Claude. (1994). "Traitement des groups nominaux en traduction automatique: opportunité d'un codage conceptuel". En: *Proceedings of the Workshop on Nominal Compounds: Multilingual Aspects of Nominal Composition*. Bouillon, Pierre; Estival, Dominique (Ed.). Université de Genève. 147-161.

L'Homme, Marie-Claude. (1997). "Méthode d'accès informatisé aux combinaisons lexicales en langue technique". En: *Meta,* 42 (1). 15-23.

Limaye, M.; Pompian, R. (1991). "Brevity versus Clarity: The Comprehensibility of Nominal Compounds in Business and Technical Prose". En: *Journal of Business Communication*, 28(1). 7-21.

Linder, Daniel. (2002). "Translating Noun Clusters and 'Nounspeak' in Specialized Computer Text". En: Chabas, José; Gaser, Rolf; Rey, Joëlle (Ed.). *Translating Science*. Barcelona: Universitat Pompeu Fabra.

López Ferrero, Carmen. (2002). Aproximación al análisis de los discursos profesionales. En: *Signos*, 35 (51-52). 195-215.

López Guix, Juan Gabriel; Minett Wilkinson, Jacqueline. (1997). *Manual de traducción español-inglés*. Barcelona: Gedisa.

Lorente, Mercè. (2001). "Altres elements lèxics". En: Solà, Joan (Dir.) *Gramàtica del català contemporani (Gcc)*. Barcelona: Empúries. 831-888.

Lorente, Mercè. (2002). Verbos y discurso especializado. *Estudios de Lingüística Española (ELIES)*, 16 [Publicación electrónica http:// elies.rediris.es]

Maalej, Zouhair. (1994). "English-Arabic Machine Translation of Nominal Compounds". En: Bouillon, P.; Estival, D. (Ed.) *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*. Geneva: ISSCO. 135-146.

Magnini, B., Strapparava, C.; Pezzulo, G.; Gliozzo, A. (2002). "The Role of Domain Information in Word Sense Disambiguation". En: *Natural Language Engineering*, 8 (4). 359–373.

Maillot, Jean. (1981). *La Traduction scientifique et technique*. Paris: Eyrolles.

Malgorzata, Tryuk. (2000). "La phraséologie en terminologie: Quelques problèmes de traduction". En: *Babel*, 46 (1). 66–76.

Maniez, Françoise. (2001). "Extraction d'une phraséologie bilingue en langue de spécialité: corpus parallèles et corpus comparable". En: *Meta*, 46 (3). 553-563.

Marcos, Francisco. (1984). *Curso de Gramática Española*. Madrid: Cincel.

Melamed, D. (1997). "A Portable Algorithm for Mapping Bitext Correspondence". En: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco, Madrid: UNED.

Méndez Cedón, Beatriz. (2002). *Estrategias fraseológicas en el género discursivo de los artículos científicos médicos en lengua inglesa*. Tesis doctoral no publicada. Universidad de Valladollid. Dir. Purificación Nistal F.

Meunier-Crespo, Mariette. (1997). "Les locutions nominales dans les dictionnaires de spécialités". En: *Meta,* 42 (1). 69-71.

Meyer, Ingrid; Mackintosh, Kristen. (1996). "Refining the Terminographer's Concept-Analysis Methods: How Can Phraseology Help?" En: *Terminology,* 3 (1). 1-26.

Miller, George 1967. *The Psychology of Communication*. New York: Basic Books.

Montero, Begoña. (1995). "Noun Premodifications vs. Postmodification in Scientific English". En: *Unesco-Alsed LSP Newsletter* 18, 2 (40). 14- 27.

Montero, Begoña. (1995). *La estructura del grupo nominal complejo en el inglés científico escrito. Sus componentes premodificadores y sus correspondencias en español.* Microfilmed Doctorate Thesis. Valencia: Universitat de Valencia.

Montero, Begoña. (1996). "Technical Communication: Complex Nominals Used to Express New Concepts in Scientific English - Causes and Ambiguity in Meaning". En: *The ESP*, 17 (1). 57-72.

Myking, Johan. (1989). "Complex Noun Phrase as a Problem of Terminological Practice". En: Laurén, Christer; Nordman, Marianne (Ed.). *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters Ltd. 265-274.

Naulleau, Eli. (1998). *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*. Tesis doctoral. Université Paris XIII. Dir. Daniel Kaiser.

Newmark, Peter. (1981). *Approaches to Translation*. London: Oxford Pergamon Press.

Newmark, Peter. (1988). *A Textbook of Translation*. London: Prentice Hall International.

Nord, Christiane. (1991). *Text Analysis in Translation*. Amsterdam-Atlanta: Rodopi.

Norman, Guy. (1999). *Cómo escribir un artículo científico en inglés*. Madrid: Astrazeneca.

Olsen, Leslie; Huckin, Thomas. (1991). *Technical Writing and Professional Communication for Nonnative Speakers of English*. Nueva York: McGraw.

Ormod, Janet. (2001). "Construction discursive de noms composés dans des textes scientifiques anglais". En: Banks, David (Ed.). *Le group nominal dans le texte spécialisé*. Paris: L'Harmattan. 9-24.

Oster, Ulrike. (2003). *Los términos de la cerámica en alemán y español: Análisis semántico orientado a la traducción de los compuestos nominales en español*. Univesitat Jaume I. Tesis doctoral. Dir. Amparo Alcina Caudet; Pilar Elena García.

Oster, Ulrike. (2005). *Las relaciones semánticas de términos polilexemáticos*. Frankfurt am Main: Peter Lang.

Palmer, Harold. (1968). *The Scientific Study and the Teaching of Languages*. Oxford: OUP.

Pinchuck, Isadore. (1977). *Scientific and Technical Translation*. London: Andre Deutsch.

Portelance, Christine. (1989). "Syntagmes et Paradigmes". En: *Meta*, 34 (3). 260-266.

Pugh, A. K.; Ulijn, Jan. M. (1984). *Reading for Professional Purposes*. London: Hienemann.

Pugh, Jeanette. (1984). "Contrastive Analysis of Noun Compound Terms in English, French, and Spanish within a Restricted, Specialized Domain". En: Hartmann, R. R. K. *Proceedings LeXeter '83*. Tübigen: Niemeyer. 395-400.

Quirk, Randoldh; Greenbaum, Sidney; Leech, Geoffrey; Svartik, Jan. (1985). *A Comprenhensive Grammar of the English Language*. London: Logman.

Quiroz, Gabriel. (2006). "Using an English-Spanish Parallel Corpus to Solve Complex Premodification in Noun Phrases". En: Gotti, M.; Sarcevic, S. (Ed.). *Translation of Specialized Text*. Linguistic Insights series. Berna: Peter Lang.

Quiroz, Gabriel. (2005a). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma*.

Papers del IULA, Sèrie Monografies, 10. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [ftp://ftp.iula.upf.es/pub/publicacions/05mon010.pdf].

Quiroz, Gabriel. (2005b). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma.* Papers del IULA, Sèrie Monografies, 10. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Quiroz, Gabriel (2005c). "Traducción de sintagmas nominales especializados extensos del inglés al español: estado de la cuestión y perspectivas". En: Rodríguez, Emma, (compiladora). *Didáctica de la traducción y la terminología* Vol. 2. Colección Estudios de Traducción y terminología. Facultad de Humanidad, Escuela Ciencias del Lenguaje, Universidad del Valle. 181-198.

Quiroz, Gabriel. (2003). *Los sintagmas nominales especializados extensos en inglés: Primera descripción en un corpus de genoma.* Trabajo de investigación de primera línea de doctorado. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona. Dir. Mercè Lorente.

Quiroz, Gabriel; De Yzaguirre, Lluís; Lorente, Mercè. (2004). "El uso de corpus paralelos para la identificación de sintagmas terminológicos extensos: ingeniería lingüística al servicio de problemas de la traducción". En: *Actas del 3.er Congreso Internacional de Traducción Especializada*. Barcelona: Universitat Pompeu Fabra.

Quiroz, Gabriel; Lorente, Merce. (2006). "Los sintagmas nominales extensos como un problema de la traducción: descripción y clasificación". En: Cabré, M-T. *et al.* (Ed.). *Actas del IX Simposio Iberoamericano de Terminología*, Riterm (Sèrie activitats, 17). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 381-392.

Quiroz, Gabriel; Lorente, Mercè; De Yzaguirre, Lluïs. (2004). “El uso de corpus paralelos para la identificación de sintagmas terminológicos extensos: ingeniería lingüística al servicio de la traducción”. En: Gaser, Rolf; Guirado, Cristina and Rey, Joël. *Insights into Scientific and Technical Translation*. Barcelona: PPU. 229-240.

Quiroz, Gabriel; Muñoz, C.; Plested, C; Giraldo, J. (2000). “Translating Medical Texts into a Foreign Language: Some Methodological Considerations”. En: *Hermes: Journal of Linguistics*, 25. 49-61.

Quiroz, Gabriel; Muñoz, Carlos. (1997). *La traducción hacia lengua extranjera mediante rastreos terminológicos en área de la malaria*. Tesis de especialización no publicada. Medellín: Universidad de Antioquia.

Rainer, Franz. (1999). “La derivación adjetival”. En: Bosque, Ignacio y Demonte, Violeta. *Gramática descriptiva de la lengua española*. Vol. 3. Madrid: Espasa. 4505-4594.

Sager, Juan Carlos. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Sager, Juan Carlos. (1992). “The Translator as a Terminologist”. En: Dollerup, Cay; Loddegaard, Anne (Ed.). *Teaching Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamins.

Sager, Juan Carlos; Dungworth, D.; McDonald, P. F. (1980). *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstteter.

Salager-Mayer, Françoise. (1984). “Compound Nominal Phrases in Scientific-Technical Literature: Proportion and Rationale”. En: Pugh, A. K.; Ulijn, J. M. (Ed.). *Reading for Professional Purposes*. London: Heinemann.

Salazar Burgos, Hada Rosabel. (2006). *Descripción y representación de los adjetivos deverbales de participio en el discurso especializado.* Tesis de DEA no publicada. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Dir. Rosa Estopà.

Scarpa, Federica. (2001). *La traduzione specializzata: Lingue speciali e mediazione lingüística*. Milano: Editore Ulrico Hoepli.

Srinivassan, V. (1993). "Developing Terminology Management Software for Translators". En: Schmitz, Klaus Dirk (Ed.). *TKE' 93*. Cologne: Indeks. 393-399.

Stockwell, Robert; Bowen, J. Donald; Martin, John W. (1965). *The Grammatical Structures of English and Spanish*. Chicago: University of Chicago Press.

Swales, John. (1974). *Writing Scientific English*. [S.L]: Nelson.

Swales, John. (1985). "The Function of One Type of Particle in a Chemistry Book". En: Trimble, Louis. *English for Science and Technology*. Cambridge: Cambridge University Press. 40-52.

Thouvenin, Susan P. (1996). *The Identification and Exemplification of Multi-Word Units within a Technical Corpus of English, Including an Investigation of Nominal Groups*. Tesis de maestría, University of Aston. [http://www.les.aston.ac.uk/lsu/diss/].

Trimble, Louis. (1985). *English for Science and Technology*. Cambridge: Cambridge University Press.

Vanderwenden, Lucretia. (1995). *The Analysis of Noun Sequences Using Semantic Information Extracted from On-Line Dictionaries*. Tesis doctoral, Georgetown University. Dir. Donald Loritz.

Varantola, Krista. (1984). *On Noun Phrase Structures in Engineering English*. Turku: Turun Yliopisto.

Varela, Soledad. (2005). *Morfología léxica: la formación de palabras*. Madrid: Gredos.

Vázquez-Ayora, Gerardo. (1977). *Introducción a la traductología*. Washington: Georgetown University Press.

Velásquez, Gonzalo. (1994). *Proceso, Métodos y Técnicas de la Traducción*. Medellín: Universidad de Antioquia.

Ventola, Eija; Mauranen, Anna. (1996). *Academic Writing Intercultural and Textual Issues*. Amsterdam: John Benjamins.

Vinay, Jean Paul; Dalbernet, Jean. (1958). *Stylistique comparée du française et de l'anglais*. Paris: Didier.

Vivaldi Palatresi, Jorge. (2004). *Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [CD-ROM] (Sèrie Tesis, 9).

Vivancos Machimbarrena, Magdalena. (1994). "Recursos estilísticos de la generalidad, impersonalidad y objetividad en el discurso científico inglés y español: su traducción". En: Charlo Brea, L. (Ed.). *Reflexiones sobre la traducción*. Cádiz: Universidad. 743-759.

Vossen, Piek (Ed.). (1998). *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic.

Walker, David G. (1993). "Translation Problems as They Occur in Everyday Practice". En: Schmitz, Klaus Dirk (Ed.). *TKE' 93*. Cologne: Indeks. 221-224.

Woolley, Reuben. (1997). *Compound Nominal Groups in the Machine Translation of Medical English: Lexical Units or Analysable Sequences?* Tesis de maestría. University of Aston. [http://www.les.aston.ac.uk/lsu/diss/].

WordNet 2.1 (2005). *Help on WordNet Terminology*. Princeton University.

Wright, Sue Ellen; Wright, Leland D. (Ed.). (1993). *Scientific and Technical Translation*. Amsterdam: John Benjamins.

Zabala, Igone. (1998). "La traducción al vasco de los sintagmas nominales complejos del lenguaje técnico". En: *Actes del III Congrés Internacional sobre Traducció, UAB*. 589-603.

Zielinski, Daniel; Ramirez, Yamile. (2005). *Research Meets Practice: T-Survey 2*005. Saarland: [S.D.]. [http://fr46.uni-saarland.de/t-survey/].

Programas de procesamiento y fuentes de consulta

Corpus de referencia del español actual (CREA) [en línea]. Madrid: Real Academia Española. [Consulta recibida el 20.02.2007]. [http://www.rae.es].

Collins COBUILD [CD-ROM]. [S.l.]: Harper Collins, 2001.

Corpus Tècnic de l'IULA [en línea]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [http://bwananet.iula.upf.edu/].

Diccionario de la lengua española [CD-ROM]. Madrid: Espasa, 2003.

Diccionario de uso del español de América y España [CD-ROM]. Barcelona: Editorial Spes –Vox, 2003.

Diccionario Espasa de Medicina [CD-ROM]. Madrid: Espasa, 1999.

Diccionario inglés-español de Ciencias de Laboratorio Clínico IFCC [en línea]. Leeds: Federación Internacional de Química Clínica, 2000. [http://www.leeds.ac.uk/ifcc/PD/dict/spandict.html].

Diccionario Mosby medicina, enfermería y ciencias de la salud inglés-español [CD-ROM]. 5ª ed. Madrid: Harcourt, 2000.

E-diccionarios Espasa [CD-ROM]. Madrid: Espasa, 2003.

EuroWordNet 1.6. [en línea]. Barcelona: Universitat Politècnica de Catalunya. [http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl o http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl].

Gran Diccionario de la Lengua Española [CD-ROM]. Barcelona: Larousse-Planeta, 1996.

IEC Multilingual Dictionary [CD-ROM]. 6ta ed. Ginebra: International Electrotechnical Commission, 2005.

IMF Terminology. Washington: International Monetary Fund, 2000. [http://www.imf.org/external/np/term/index.asp?index=eng&index_langid=1]

ISI Multilingual Glossary of Statistical Terms [en línea]. La Haya: International Statistical Institut, 2006. [http://isi.cbs.nl/glossary/index.htm]

Lluís de Yzaguirre. (2004). *Repoker: programa para la extracción de datos lingüísticos etiquetados*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Longman Dictionary of Contemporary English [CD-ROM]. Essex: Pearson Education Limited, 2003.

Machinese Phrase Tagger online demo. Helsinki: Connexor Oy, 2005. [www.conexor.eu].

Random House Webster's Unabridged Dictionary [CD-ROM]. [S.L]: Random House Reference, 2006.

Reed, Alan. (2002). *Simple Concordance Programme*, 4.0.7 [programa]. [http://www.textworld.com/scp/index.html]

Routledge Spanish Dictionary of Business, Commerce and Finance/Diccionario Inglés de Negocios, Comercio y Finanzas [CD-ROM]. London: Routledge, 1999.

Routledge Spanish Technical Dictionary/Diccionario técnico inglés [CD-ROM]. London: Routledge, 1998.

Stedman's Medical Dictionary 3.0 [CD-ROM]. [S.L]: Baltimore Williams & Wilkins, 1996.

UMLS Knowledge Source Server (UMLSKS) [en línea]. Washington: National Library of Medicine, 2006. [http://umlsks.nlm.nih.gov].

WordNet 2.1. [programa]. Princeton: Universidad de Princeton, 2006. [http://wordnet.princeton.edu].

Anexo 1: Listado de patrones de extracción en inglés

Nº del patrón	N.º de tokens	Patrón simple	Ejemplo	Patrón IULA
1	3	Adv PP Adj N	genetically controlled in vivo tests / genetically determined immune response	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="JA"] [pos="NN.*"]
2	3	Adv PP N	oxidatively damaged dna / chromosomally encoded efflux	[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"]
3	4	Adv PP N N	electrophoretically altered migration patterns / exogenously added mrna molecules	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"] [pos="NN.*"]
4	5	Adv PP N N N	highly conserved tyrosine kinase phosphorylation / covalently closed plasmid dna band	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]

5	4	Adv PP X N	chromosomally encoded penicillin-resistance genes / highly activated myofibroblastic cells	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
6	4	Adv Adj Adj N	morphologically identifiable apoptotic cells / right ventricular free wall	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="JA"] [pos="JA"] [pos="NN.*"]
7	3	Adv Adj N	right ventricular myocardium / sexually dimorphic cell	[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="JA"] [pos="NN.*"]
8	4	Adv Adj N N	in situ squamous carcinoma cells / clinically x-linked ichthyosis patients	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="JA"] [pos="NN.*"] [pos="NN.*"]
9	3	Adv PP N	concomitantly reexpressed albumin / radioactively labelled nucleotides	a:[pos="D6" & lemma=".*ly|in vitro|in vivo|ex vivo|very|long|overall|well|rather|right|in situ|upstream|a priori|almost|already|somewhat"] [pos="V6A66"&lemma!="have|be|do|make|suggest|reside"] [pos="NN.*"]
10	3	PP N N	pulverized rat chow / verified mutation carriers	[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"] [pos="NN.*"]
11	3	PP Adj N	polarized epithelial cells / blocking repetitive hybridization	[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="JA"] [pos="NN.*"]{1,2}
12	4	PP Adj N N	acquired mitochondrial dna mutations / circulating monoclonal ig protein	a:[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="JA"] [pos="NN.*"] [pos="NN.*"]

13	4	PP N N N	pulsed field gel electrophoresis / conserved tyrosine kinase phosphorylation	a:[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
14	3	PP|PPi PP N	corresponding cloned cdna / remaining labeled material	a:[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="V6A.*"&lemma!="have|be|do|make|suggest|reside"] [pos="NN.*"]
15	3	PP N N	activated huvec rna / advanced yac library	[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
16	3	Adj PP N	open reading frame / putative coding region	[pos="JA"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"]
17	3	Adj Adj N	human fetal brain / neuronal nicotinic acetylcholine	[pos="JA"] [pos="JA"] [pos="NN.*"]{1,4}
18	4	Adj Adj Adj N	human mitochondrial ribosomal protein / human acute lymphoblastic leukemia	a:[pos="JA"] [pos="JA"] [pos="JA"] [pos="NN.*"]
19	5	Adj Adj Adj N N	human mitochondrial ribosomal protein genes / dominant nocturnal frontal lobe epilepsy	a:[pos="JA"] [pos="JA"] [pos="JA"] [pos="NN.*"] [pos="NN.*"]

20	4	Adj Adj N N	somatic mitochondrial dna mutations / neuronal nicotinic acetylcholine receptors	[pos="JA"] [pos="JA"] [pos="NN.*"] [pos="NN.*"]
21	5	Adj Adj N N N	human mitochondrial atp-binding cassette membrane / high mtdna-specific pcr band intensities	a:[pos="JA"] [pos="JA"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
22	4	Adj N Adj N	common c57bl/6j inbred background / tight west african cluster	a:[pos="JA"] [pos="NN.*"] [pos="JA"] [pos="NN.*"]
23	3	Adj N N	prandial insulin infusions / wild-type core protein	[pos="JA"] [pos="NN.*"] [pos="NN.*"]
24	4	Adj N N N	fetal brain cdna library / human r-banded metaphase chromosomes	[pos="JA"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
25	5	Adj N N N N	intracellular tyrosine kinase phosphorylation motif / histiocytic lymphoma cell line u937	a:[pos="JA"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
26	4	Adj N Adj N	human apoe genomic dna / fractional sex-average genetic map	a:[pos="JA"] [pos="X"&lemma!="that|which|who"] [pos="JA"] [pos="NN.*"]

27	3	Adj N N	high mannose oligosaccharides / neuronal apoe immunoreactivity	[pos="JA"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
28	4	Adj N N N	native agarose gel electrophoresis / allele-specific oligonucleotide analysis hybridization	a:[pos="JA"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"] [pos="NN.*"]
29	3	D|X X N	seventh transmembrane domains / xxx egfr mutant	a:[pos="MO6"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
30	4	N Adv N N	fluorescence in situ hybridization experiments / fluorescence in situ hybridization probes	a:[pos="NN.*"] [pos="D6"] [pos="NN.*"] [pos="NN.*"]
31	4	N PP Adj N	dna binding regulatory proteins / membrane bound intracellular organelles	a:[pos="NN.*"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="JA"] [pos="NN.*"]
32	3	N PP N	dna sequencing kit / calcium sensing receptor	[pos="NN.*"] [pos="H6"&word!="containing|including|having|containing|producing|uing|causing|identifying|involving"] [pos="NN.*"]
33	4	N PP N N	cascade mediating egfr-induced mitogenesis / shaker related subfamily member	a:[pos="NN.*"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"] [pos="NN.*"]

34	4	N Adj Adj N	female mammalian somatic cells / plasma free fatty acid	a:[pos="NN.*"] [pos="JA"] [pos="JA"] [pos="NN.*"]
35	3	N Adj N	immunoglobulin heavy chain / muc7 genomic clones	[pos="NN.*"] [pos="JA"] [pos="NN.*"]
36	4	N Adj N N	moloney murine leukemia virus / abi373a automatic dna sequencer	a:[pos="NN.*"] [pos="JA"] [pos="NN.*"] [pos="NN.*"]
37	4	N Adj N N	baseline concurrent anti-hiv-1 treatment / surface major histocompatibility complexes	a:[pos="NN.*"] [pos="JA"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
38	4	N N PP N	terminator cycle sequencing kit / arabidopsis suspension cultured cells	a:[pos="NN.*"] [pos="NN.*"] [pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="NN.*"]
39	3	N N N	polymerase chain reaction / plasmid dna purification	[pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
40	4	N N N N	restriction fragment length polymorphism / potassium channel gene cluster	[pos="NN.*"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]
41	5	N N N N N	brain cdna lambda zap library / egfr-ras-map kinase signal transduction pathway	a:[pos="NN.*"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"] [pos="NN.*"]

42	3	N N|X N	amino acid sequence / transmission disequilibrium test	[pos="NN.*"] [pos="NN.*|X"&lemma!="that|which|who"] [pos="V6R6S"&lemma!="have|be|do|make|suggest|reside"]
43	3	N X N	restriction endonuclease digestion / ct box-binding proteins	[pos="NN.*"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
44	4	N X N N	drosophila dlg tumor suppressor / fibrosis transmembrane conductance regulator	a:[pos="NN.*"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"] [pos="NN.*"]
45	3	PP X N	primed first-strand cdna / generalized tonic-clonic seizures	a:[pos="V6A66"&lemma!="have|be|do|make|suggest|reside"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
46	3	X Adj N	apoe transgenic mice / 12-specific genomic library	[pos="X"&lemma!="that|which|who"] [pos="JA"] [pos="NN.*"]
47	3	X N N	agarose gel electrophoresis / polyacrylamide gel electrophoresis	[pos="X"&lemma!="that|which|who"] [pos="NN.*"] [pos="NN.*"]
48	3	X X N	laser-desorption time-of-flight mass / calcium-modulating cyclophilin ligand	[pos="X"&lemma!="that|which|who"] [pos="X"&lemma!="that|which|who"] [pos="NN.*"]
49	3	Adj Adj N	bilateral central epileptiform / experimental autoimmune uveoretinitis	a:[pos="JA"] [pos="JA"] [pos="X"&lemma!="that|which|who"]

50	4	PP Adj Adj N	polarized renal epithelial cells / targeted green fluorescent protein	a:[pos="H6"&word!="containing|including|having|containing|producing|using|causing|identifying|involving"] [pos="JA"] [pos="JA"] [pos="NN.*"]

Anexo 2: Listado de patrones de extracción en español

N.º del patrón	N.º de tokens	Patrón simple	Ejemplo	Patrón IULA
1	4	Adj N Prep D N Prep D N	escasa especificación de la localización de algunas poblaciones	[pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="E.*"] [pos="N5.*"]
2	3	Adj N Prep N	diferentes enzimas de restricción	[pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
3	4	N Adj Prep N Prep N	síndrome dismetabólico de sobrecarga de hierro	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
4	5	N Adj Prep N Prep N Adj	afectación pulmonar en forma de neumonía atípica	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
5	5	N Adj Prep N Prep N Prep N	ND	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
6	3	Adj Prep N Prep N	síndrome dismetabólico de sobrecarga de hierro	[pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
7	3	Adj N Adj	alto peso molecular	[pos="JQ.*"] [pos="V.*"&lemma!="ser|parecer|representar|dejar|tener|incluir presentar|evolucionar|representar|producir|encontrar|revelar|determinar|distribuir|situar|quedar|permanecer|admitir|utilizar|conseguir|observar|amplificar|analizar|mostrar|describir|expresar|guardar"] [pos="JQ.*"]
8	3	Adj N Adj	ciertos compuestos organoclorados	[pos="JQ.*"] [pos="H.*" & lemma!="ser|parecer|representar|dejar|tener|incluir presentar|evolucionar|representar|producir|encontrar|revelar|determinar|distribuir|situar|quedar|permanecer|admitir|utilizar|conseguir|observar|amplificar|analizar|mostrar|describir|expresar|guardar"] [pos="JQ.*"]
9	3	Adj PP Prep N	demarcadores ligados a loci	[pos="JQ.*"] [pos="VC.*"] [pos="P"] [pos="N5.*"]

10	4	N Adj PP Prep N	células mutadoras activadas por linfocitos	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*"] [pos="P"] [pos="N5.*"]
11	3	N Adj PP	muerte celular programada	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"]
12	5	N Adj PP Prep N Adj	anticuerpos monoclonales ligados a partículas magnéticas	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
13	4	N Adj PP Prep N	secuencias cortas repetidas en tándem	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"] [pos="P"] [pos="N5.*"]
14	5	N Adj PP Prep D N Adj	genes supresores relacionados con la neoplasia vesical	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"]
15	4	N Adj PP Prep D N	miopatía miotubular ligada al cromosoma	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"] [pos="P"] [pos="A.*"] [pos="N5.*"]
16	3	N Adv Adj	loci altamente polimórficos	[pos="N5.*"] [pos="D6"] [pos="JQ.*"]
17	4	N PP Adj Prep D N	ND	[pos="N5.*"] [pos="H.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]
18	4	N Adj Conj Adj Adj	cambios genotípicos y fenotípicos consecuentes	[pos="N5.*"] [pos="JQ.*"] [pos="C"] [pos="JQ.*"] [pos="JQ.*"]
19	5	N Adj Adv PP Prep D N	ND	[pos="N5.*"] [pos="JQ.*"] [pos="D4"] [pos="VC.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]
20	4	N Adj Adv Adj	cáncer vesical cistoscópicamente visible	[pos="N5.*"] [pos="JQ.*"] [pos="D6"] [pos="JQ.*"]
21	3	N Adj PP	genotipo heterocigoto compuesto	[pos="N5.*"] [pos="JQ.*"] [pos="H.*|VC.*" & lemma!="asociar|dar|deber|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"]

22	3	N Adj D Adj	diabetes mellitus insulinodependiente	[pos="N5.*"] [pos="JQ.*" & word!="afecta"] [pos="JQ.*" & word!="afecta"]
23	4	N Adj Adj Adj	patrón mendeliano autosómico dominante	[pos="N5.*"] [pos="JQ.*"] [pos="JQ.*"] [pos="JQ.*"]
24	4	N Adj Adj N	hepatitis vírica crónica b	[pos="N5.*"] [pos="JQ.*"] [pos="JQ.*"] [pos="N5.*"]
25	4	N Adj Adj Prep D N	anormalidades genéticas responsables de la tumorigénesis	[pos="N5.*"] [pos="JQ.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N5-F6"]
26	3	N Adj N	óxido nítrico sintasa	[pos="N5.*"] [pos="JQ.*" & lemma!="rara|vez"] [pos="N5.*" & lemma!="rara|vez"]
27	4	N Adj Prep D Adj N	surco mayor de la doble hélice	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="JQ.*"] [pos="N5.*"]
28	5	N Adj Prep D Adj N Adj	ND	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="JQ.*"] [pos="N5.*"] [pos="JQ.*"]
29	3	N Adj Prep D N	brazo corto del cromosoma	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]
30	4	N Adj Prep D N Adj	asta anterior de la médula espinal	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"]
31	4	N Adj Prep D N Prep N	aislamiento selectivo mediante la identificación de islas	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
32	4	N Adj Prep Adj N	tumores vesicales de alto grado	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"]
33	5	N Adj Prep Adj N Prep N	estructura espacial de varios dedos de zinc	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
34	5	N Adj Prep N Prep N Prep N	visualización directa tras tinción con bromuro de etidio	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N.*"] [pos="P"] [pos="N.*"] [pos="P"] [pos="N5.*"]
35	3	N Adj Prep N	fenotipo mutador de microsatélites	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"]
36	5	N Adj Prep N Adv Adj	ND	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="D6"] [pos="JQ.*"]
37	4	N Adj Prep N Adj	alelo largo sin secuencias flanqueantes	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
38	4	N Adj Prep N N	terapias regenerativas con células madre	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="N.*"]
39	4	N Adj Prep N Prep D N	actividades enzimáticas de miofosforilasa en la biopsia	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]

40	5	N Adj Prep N Prep D N Adj	acidificación intracelular en respuesta al ejercicio isquémico	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"]
41	4	N Adj Prep N Prep N	distancias genéticas entre los diversos grupos étnicos	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
42	5	N Adj Prep N Prep N Prep N	defectos moleculares en pacientes con déficit en miofosforilasa	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
43	5	N Adj Prep N Prep N Adj	episodios tromboembólicos por afectación de vasos pequeños	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
44	4	N Adj Prep N PP	clonaje posicional de genes mutados	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="VC.*"]
45	3	N Adj PP	genes supresores relacionados con la neoplasia vesical	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"]
46	4	N Adj PP Adv	enfermedad coronaria demostrada angiográficamente	[pos="N5.*"] [pos="JQ.*"] [pos="VC.*" & lemma!="deber|asociar|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"] [pos="D6"]
47	3	N N Adj	células madre embrionarias	[pos="N5.*"] [pos="N5.*"] [pos="JQ.*"]
48	3	N N N	proteína quinasa c	[pos="N5.*"] [pos="N5.*"] [pos="N5.*"]
49	3	N N Prep N	hibridación in situ con fluorescencia	[pos="N5.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
50	4	N N Prep N Adj	fertilización in vitro con transferencia embrionaria	[pos="N5.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
51	4	N N Prep N Prep N	actividad tirosincinasa sin necesidad de dimerización	[pos="N5.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
52	5	N N Prep N Prep N Adj	trisomía x con genes de crecimiento activos	[pos="N5.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
53	3	N Prep D N PP	citólisis de las células infectadas	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N.*"] [pos="H.*" & lemma!="asociar|dar|deber|administrar|encontrar|colocar|situar|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"]

54	3	N Prep D N Prep D N	intrón del gen de la mioglobina	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N.*"] [pos="P"] [pos="A.*"] [pos="N.*"]
55	3	N Prep D N Adj	betalactamasas de las bacterias gramnegativas	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"]
56	5	N Prep D N Adj Prep D N Adj	motoneuronas del asta anterior de la médula espinal	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N.*"] [pos="JQ.*"]
57	6	N Prep D N Adj Prep D N Prep N Adj	impacto de los fármacos antihipertensivos sobre las cifras de colesterol plasmático	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"]
58	4	N Prep D N Adj Prep D N	motoneuronas del asta anterior de la médula	[pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N.*"]
59	3	N Prep Adj Conj Adj	tratamiento con anticolinérgicos y anti-inflamatorios	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="C"] [pos="JQ.*"]
60	3	N Prep Adj Adj	técnicas de genética molecular	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="JQ.*"]
61	4	N Prep Adj Adj Prep N	utilización de diversas técnicas de clonaje	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="JQ.*"] [pos="P"] [pos="N.*"]
62	5	N Prep Adj Adj Prep N Adj	introducción de nuevas técnicas de biología molecular	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="JQ.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"]
63	4	N Prep Adj Adj Adj	neoplasias con diferente potencial biológico	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="JQ.*"] [pos="JQ.*"]
64	3	N Prep Adj N	lipoproteínas de alta densidad	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"]
65	4	N Prep Adj N Adj	técnicas de alta resolución cromosómica	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"] [pos="JQ.*"]
66	4	N Prep Adj N Prep N	haplotipos con fuertes desequilibrios de ligamiento	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
67	5	N Prep Adj N Prep N Adj	activación de diferentes rutas de transmisión intracelular	[pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
68	3	N Prep N Adj	agenesia de cuerpo calloso	[pos="N5.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"]
69	4	N Prep N Adj Adj	patrón de herencia autosómico dominante	[pos="N5.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"] [pos="JQ.*"]
70	4	N Prep N Adv Adj	azoospermia en varones sexualmente maduros	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="D6"] [pos="JQ.*"]
71	3	N Prep N PP	tinción con anticuerpos conjugados	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="H.*" & lemma!="asociar|dar|deber|administrar|encontrar|colocar|situa

				r|ayudar|iniciar|derivar|conocer|ocurrir|dar|liberar|contener|involucrar|denominar|alcanzar"]
72	4	N Prep N Adj Prep D N	migración de genes mitocondriales al núcleo	[pos="N5.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"] [pos="P"] [pos="A.*"] [pos="N.*"]
73	4	N Prep N Adj Prep N	carcinoma de células transicionales de vejiga	[pos="N5.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"] [pos="P"] [pos="N.*"]
74	5	N Prep N Prep D N Prep N Prep N	detección de polimorfismos mediante la introducción de sitios de restricción	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
75	5	N Prep N Prep D N Prep N Prep D N	problemas de especificidad de la técnica de detección de la mutación	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]
76	4	N Prep N Prep D N Prep N	detección de polimorfismos mediante la introducción de sitios	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
77	5	N Prep N Prep D N Prep N Adj	reacción en cadena de la polimerasa con retro-transcripción previa	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
78	5	N Prep N Prep Adj N Prep N	presencia de multitud de finas gotas de ésteres	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="JQ.*"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
79	3	N Prep N Prep N	digestión con enzimas de restricción	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N.*"]
80	5	N Prep N Prep N Adv Adj	confirmación de azoospermia en varones sexualmente maduros	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="D6"] [pos="JQ.*"]
81	4	N Prep N Prep N Adj	diferenciación de células con potencial adipogénico	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
82	4	N Prep N Prep N Prep D N	transversión de guanina a citosina en el nucleótido	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="A.*"] [pos="N5.*"]
83	4	N Prep N Prep N Prep N	exceso de secreción de hormona de crecimiento	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
84	4	N Prep N Prep N PP	estrato de fibroblastos de ratón irradiados	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="VC.*"]
85	4	N Prep N PP Adv	tipos de filamentos orientados horizontalmente	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="VC.*"] [pos="D6"]

86	3	N Prep N X	diabetes de tipo 1	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="X"]
87	3	N Prep N Adj	agenesia de cuerpo calloso	[pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
88	3	N PP Adj	agua destilada estéril	[pos="N5.*"] [pos="VC.*"] [pos="JQ.*"]
89	4	N PP Prep D N Adj	alteraciones detectadas en la rm convencional	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="A.*"] [pos="N.*"] [pos="JQ.*"]
90	5	N PP Prep N Adj Adj	copias duplicadas por selección natural positiva	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"] [pos="JQ.*"]
91	4	N PP Prep N Adj	deficiencia combinada de hormonas hipofisarias	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"]
92	5	N PP Prep N Adj Prep N	estudios realizados con pautas cortas de doxiciclina	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"]
93	3	N PP Prep N	secuencias repetidas en tándem	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N5.*"]
94	5	N PP Prep N Prep N Adj	loci detectados por sondas de locus específico	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
95	4	N PP Prep N Prep N	microsatélite constituido por repeticiones en tándem	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
96	3	N PP Prep W	conejos inmunizados con grf-1	[pos="N5.*"] [pos="VC.*"] [pos="P"] [pos="W"]
97	4	N Adj Prep N Prep N	síndrome dismetabólico de sobrecarga de hierro	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]
98	5	N Adj Prep N Prep N Adj	afecciones otorrinolaringológicas en pacientes con retinosis pigmentaria	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
99	5	N Adj Prep N Prep N Prep N	adsorción diferencial mediante células de riñón de cobaya	[pos="N5.*"] [pos="JQ.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"] [pos="P"] [pos="N5.*"]

Anexo 3: Listado de patrones de finales en inglés

Patrón	Ejemplo	Tokens	Frecuencia
N N N	polymerase chain reaction	3	317
Adj N N	horizontal gene transfer	3	254
Adj Adj N	human genomic DNA	3	113
N Adj N	platelet dense granules	3	62
PP N N	reduced insulin responsiveness	3	51
Adj N N N	fetal brain cDNA library	4	33
PP Adj N	polarized epithelial cells	3	31
N N N N	restriction fragment length polymorphism	4	28
Adv Adj N	anatomically modern humans	3	25
N PP N	ATP binding site	3	23
Adj Adj N N	human peripheral blood lymphocytes	4	16
Adv PP N	genetically engineered microorganisms	3	19
Adj PP N	neutral buffered formalin	3	13
N Adj N N	immunoglobulin heavy chain locus	4	10
Adj N Adj N	human APOE genomic DNA	4	9
PP N N N	pulsed field gel electrophoresis	4	6
PP Adj N N	inherited mitochondrial DNA diseases	4	5
Adj Adj Adj N	total human genomic DNA	4	4
Adv Adj N N	highly deleterious mtDNA mutations	4	4
Adv PP N N	highly conserved phosphotyrosine domain	4	4
Adj Adj N N N	human fetal brain cDNA library	5	3
Adj N N N N	fat cell size distribution profile	5	3
Adv Adj Adj N	morphologically identifiable apoptotic cells	4	3

N N N N N	V1aR mRNA transcription start site	5	3
N N PP N	Arabidopsis suspension cultured cell	4	3
PP Adj Adj N	polarized renal epithelial cells	4	3
Adv PP Adj N	genetically determined immune response	4	2
N Adj Adj N	GEM11 human genomic library	4	2
N PP N N	double stranded plasmid DNA	4	2
Adj Adj Adj N N	mature neuronal nicotinic acetylcholine receptors	5	1
Adv PP N N N	covalently closed plasmid DNA band	5	1
N Adv N N	fluorescence in situ hybridization probes	4	1
PP PP N	written informed consent	3	1

Anexo 4: Listado de patrones de finales en español

Patrón	Ejemplo	Frecuencia	Porcentaje
N Prep N Adj	virus de la inmunodeficiencia humana	343	31,66
N Adj Prep N	artrosis degenerativa de la columna	175	16,13
N Prep N Prep N	electroforesis en gel de agarosa	118	10,68
N Adj Adj	diabetes mellitus insulinodependiente	73	6,73
N Adj PP	células alveolares descamadas	53	4,88
N Adj Prep N Adj	membrana apical de las células epiteliales	40	3,68
N Prep N Adj Prep N	constricción de las arterias coronarias del corazón	36	3,31
Adj N Prep N	alto grado de polimorfismo	30	2,76
N Adj Prep N Prep N	secreción excesiva de hormona de crecimiento	29	2,67
N Prep Adj N	sulfonilurea de alta afinidad	20	1,84
N PP Prep N	oligonucleótidos repetidos en tándem	14	1,29
N Prep N Prep N Prep N	hipocrecimiento por anomalías en genes de los gonosomas	12	1,11
N Prep N Prep N Adj	electroforesis en geles de campos pulsantes	11	1,01
N Prep N PP	hibridación con sonda marcada	10	0,93
N Prep N Adj Adj	inoculación con adenopatías satélites axilares	9	0,83
N Adj PP Prep N	proteína mitocondrial sintetizada en el citosol	8	0,74
N N Adj	hormona somatomamotropina coriónica	8	0,74
N Adj Prep N Prep N Adj	síndrome dismetabólico de sobrecarga de hierro heterocigoto	7	0,64
N Adv Adj	loci altamente polimórficos	7	0,65
N Adj N	hepatitis vírica C	6	0,55
N Adj Prep Adj N	cromatografía líquida de alta resolución	6	0,55
Adj N Adj	alto peso molecular	5	0,46
N N Prep N	amfotericina B en liposomas	5	0,46

N Adj Prep N Prep N Prep N	visualización directa tras tinción con bromuro de etidio	4	0,37
N Adj Adj Adj	poliquistosis renal autosómica recesiva	3	0,28
N Adj Prep N N	terapias regenerativas con células madre	3	0,28
N PP Prep N Adj	metilasas codificadas por los genes kgmA	3	0,56
N Prep Adj N Adj	Hibridación con oligonucleótidos alelo específicos	3	0,28
N Prep N Adj Prep N Adj	motoneuronas del asta anterior de la médula espinal	3	0,28
N Adj PP Prep N Adj	anticuerpos monoclonales ligados a partículas magnéticas	2	0,18
N N N	citocromo c oxidasa	2	0,18
N PP Prep N Prep N	lactamasas codificadas en plásmidos de enterobacterias	2	0,18
N Prep Adj N Prep N	resistencia a diferentes clases de antibióticos	2	0,18
N Prep N N	diabetes de tipo 1	2	0,18
N Prep N Prep N Prep N Prep N	método de delección del cúmulo de hierro en el cuerpo	2	0,18
Adj N Prep N Prep N	escasa especificación de la localización de algunas poblaciones	1	0,09
N Adj Adj N	Hepatitis vírica crónica B	1	0,09
N Adj Adj Prep N	anormalidades genéticas responsables de la tumorigénesis	1	0,09
N Adj Adv Adj	cáncer vesical cistoscópicamente visible	1	0,09
N Adj Adv PP Prep N	enfermedad neuromuscular no ligada al sexo	1	0,09
N Adj PP Adv	bacterias gramnegativas relacionadas serológicamente	1	0,09
N Adj Prep Adj N Adj	valores predictivos de los diversos métodos diagnósticos	1	0,09
N Adj Prep Adj N Prep N	genoma humano con idéntico mapa de restricción	1	0,09
N Adj Prep N Adv Adj	fuentes idóneas de linfocitos inmunológicamente activos	1	0,09
N Adj Prep N PP	clonaje posicional de genes mutados	1	0,09
N N Prep N Adj	actividad transferasa en vellosidades curiales	1	0,09
N N Prep N Prep N	actividad proteincinasa sobre residuos de tirosina	1	0,09
N N Prep N Prep N Adj	trisomía X con genes de crecimiento activos	1	0,09
N PP Adj	agua destilada estéril	1	0,09
N PP Adj Prep N	alelos clonados diferentes del locus	1	0,09
N PP Prep N Prep N Adj	ratas modificadas por medio de ingeniería genética	1	0,09
N Prep Adj Adj Prep N	cultivos con medios pobres en folato	1	0,09
N Prep Adj N Prep N Adj	endarteritis de pequeños vasos con proliferación endotelial	1	0,09
N Prep N Adj Prep N Prep N Adj	actividad de la enzima responsable de la síntesis de óxido nítrico	1	0,09

N Prep N Adv Adj	azoospermia en varones sexualmente maduros	1	0,09
N Prep N PP Adv	familia de secuencias relacionadas evolutivamente	1	0,09
N Prep N Prep Adj N Prep N	lugares de reconocimiento para distintos factores de transcripción	1	0,09
N Prep N Prep N Adv Adj	confirmación de azoospermia en varones sexualmente maduros	1	0,09
N Prep N Prep N PP	hemoperfusión con cartucho de carbón activado	1	0,09
N Prep N Prep N Prep N Adj	vía de transmisión de la señal de modo constante	1	0,09