8. Esquemas de representación para el análisis extensivo de la cohesión

El hecho de que la mayoría de los textos tengan una longitud incompatible con el análisis intensivo o con la metodología de la lingüística de corpus nos ha impulsado a diseñar una dinámica de trabajo diferente para el *análisis extensivo de la cohesión*. Se contará con el soporte de cuatro elementos:

(1) El programa de análisis léxico *Wordsmith Tools*
(2) El generador de cadenas léxicas *Hesperus* (véase pp. 134-138)
(3) La hoja de cálculo *Microsoft Excel*
(4) El analisis del investigador, que aporta su conocimiento de la lengua y lo complementa con la información que proporciona el experto en oncología.

8.1. Descripción de los resultados obtenidos con el programa informático *Hesperus*

Para cada texto que sirve de *input* a *Hesperus*, este programa confecciona un conjunto de documentos en HTML con los siguientes elementos:

a) perspectiva general (*Overview*)
b) perfil genérico del documento (*Generic document profile*)
c) versión del texto completo donde se destacan las distintas cadenas léxicas reconocidas
d) representación de cada cadena con todos los lexemas que la componen y el tipo de relación entre ellos (*Chains*).

Para explicar en qué consiste cada uno, hemos seleccionado el *output* a partir de un texto para profesionales de la salud sobre el cáncer no microcítico de pulmón (QDT2) (Cfr. Apéndice II, texto 10).

8.1.1. Perspectiva general

En la *perspectiva general*, aparecen todos los textos analizados y se les atribuye un número con dos decimales que mide la similitud de los mismos con respecto al texto

que *Hesperus* ha seleccionado como prototipo, el texto QDT1 (Cfr. Apéndice II, texto9). Para los textos QDT1 y QDT2, aparecería en pantalla lo siguiente[106]:

TABLA 29: Perspectiva general correspondiente a los textos QDT1 Y QDT2.

Case: 35
QDT1.txt 1.00 (505 words, 10536 total)
(Chains,Profile)

Case: 34
QDT2.txt 0.20 (2470 words, 46217 total)
(Chains,Profile)

El número 0.20 que sigue a QDT2.txt indica el grado de similitud de ese texto con respecto a QDT1. A este texto prototípico, se le atribuye un valor de 1.00, ya que es similar a sí mismo. El número 2470 representa el número de palabras plenas del texto, porque, como vimos en la p. 136, *Hesperus* elimina en el análisis las *palabras forma*, ya que no son susceptibles de crear cohesión léxica.

En el mismo paréntesis, después del 2470, un número (*raw score*) representa la suma de todos los valores que *Hesperus* otorga a las categorías conceptuales del *Roget* activadas. Para visualizarlas sólo hay que hacer click en el perfil del documento (*profile*). Este valor no indica por sí sólo la calidad cohesiva del texto dado que, a mayor longitud del texto, se obtiene una puntuación mayor. Por este motivo, se relacionará este valor con el número total de palabras. En el perfil del documento aparece desglosado este valor.

Desde el perfil también podemos visualizar la versión del texto (QDT1.txt, QDT2.txt) que contiene hipervínculos hacia las cadenas léxicas (Chains).

8.1.2. Perfil del documento

El *perfil del documento* incluye las categorías conceptuales del tesauro que se activan gracias a la cohesión que construyen las palabras plenas del texto. Como este perfil es demasiado extenso, fragmentamos el perfil completo y sólo incluimos los conceptos

[106] La perspectiva general completa aparece en el Apéndice VIIa.

más representativos y los menos representativos. Editamos el perfil completo en el apéndice VIIb.

TABLA 30: Fragmento del perfil del texto QDT2.

NON-SMALL CELL LUNG CANCER

```
Document Profile: No file name (default), 196 Features, Total
46217 (0)

   therapy_658_4268_n Percent: 21.07,              Value: 9737
   sick-person_651_4202_n Percent: 7.02,           Value: 3246
   show_522_3381_v Percent: 5.69,                  Value: 2628
   minuteness_196_1271_n Percent: 3.46,            Value: 1600
   surgery_658_4267_n Percent: 3.02,               Value: 1396
   medical_658_4274_a Percent: 2.72,               Value: 1255
   experiment_461_2959_n Percent: 2.64,            Value: 1220
   cancer_651_4195_n Percent: 2.56,                Value: 1183
   respiratory-disease_651_4192_n Percent: 1.90,   Value: 880

    [...]

   modality_7_44_n Percent: 0.11,                  Value: 50
   lasting_113_699_a Percent: 0.11,                Value: 50
   cardiovascular-disease_651_4193_n Percent: 0.11Value: 50
   unprosperous_731_4742_a Percent: 0.11,          Value: 50

 Total Percent: 97.47
```

Antes del listado de conceptos, se especifica la cantidad de categorías conceptuales que subyacen las 2470 palabras plenas del texto. A esto hace alusión la referencia *196 Features.* No obstante, de estas 196 categorías, sólo aparecen en el perfil aquellas que contribuyen más de un 0,11 % a la cohesión del texto. Le sigue el ya mencionado *raw score* del documento (46217).

A continuación, cada categoría lleva una referencia a la jerarquía del *Roget's Thesaurus* (_658_4268) y a la categoría gramatical (nombre, verbo, adjetivo o adverbio) asociada a la misma. Le sigue un porcentaje que se obtiene de dividir el *raw score* del texto entre el valor asignado a esa categoría conceptual. Por ejemplo, el concepto principal del texto, THERAPY, pertenece a la categoría 658 (IMPROVEMENT), se actualiza mediante un sustantivo y representa un 21,07% de los conceptos activados en el texto. Confirmaremos este porcentaje cuando veamos que el concepto TREATMENT es más relevante en el texto analizado.

8.1.3. Versión del texto con hipervínculos hacia las cadenas léxicas

Hesperus también confecciona una versión del texto en la que las palabras pertenecientes a la misma cadena léxica aparecen con el mismo color (véase Apéndice VIIc). Presentamos el primer párrafo del texto en cuestión, donde las palabras que introducen cinco cadenas léxicas están subrayadas, en color azul y llevan un superíndice: *small, lung-cancer, cure, produce* y *partial.* Estos lexemas inician las cadenas que se hilan, respectivamente, a partir de los conceptos: MINUTENESS, PATHOLOGY OF LUNG CANCER, TREATMENT, CAUSE y COMPOSITION.

TABLA 31: Versión HTML del texto QDT2.

---------- NON- small[3] *cell* lung-cancer[1] ----------
********** GENERAL INFORMATION
Non- **small cell lung-cancer** (NSCLC) is a heterogeneous aggregate of at least three distinct histologies of **lung-cancer** including epidermoid or squamous *carcinoma*, adenocarcinoma, and large **cell carcinoma**. These histologies are often classified together because, when localized, all have the potential for cure[0] with *surgical* resection. Systemic chemotherapy can produce[2] objective partial[13] *responses* and palliation of *symptoms* for short durations. Local control can be achieved with radiation in a large number of *patients* with unresectable *disease*, but **cure** is seen only in a **small** *minority* of **patients**.

Como vemos en el texto, el primer elemento de cada cadena lleva un superíndice con un número cuyo color indica que todas las palabras con tipografía en ese color van a encajar en esa cadena. Y así, en la primera línea, el término *lung cancer* va acompañado de un número en verde azulado. Todos los lexemas del texto con este color [*lung cancer* (2), *carcinoma* (2), *symptoms*, *patients* (2) y *disease*] están relacionados con *lung cancer* (primera línea) y forman una cadena, a la que se puede acceder mediante un vínculo de hipertexto. Según la tabla 31, pertenecen a esta cadena los lexemas *lung cancer*, que aparece en tres ocasiones, *carcinoma* y *patients*, que se repiten en dos ocasiones, *symptoms* y *disease*.

La tipografía de los lexemas "encadenados" es informativa del tipo de relación semántica que se da, de acuerdo con las siguientes convenciones:

- **identity** (negrita): relación entre formas idénticas o entre sustantivos que cambian sólo respecto al número. Ejemplo: *patient — patients*.
- category (subrayado): relación entre lexemas que pertenecen a una misma categoría del tesauro *Roget.* Ejemplo: *diagnosis – prognosis*, asociados a la categoría pathology_651_4203_n.
- *group* (cursiva): relación entre lexemas pertenecientes a categorías vecinas dentro de un mismo *grupo de categorías* del tesauro. Ejemplo: *lung cancer – carcinoma*, relacionados respectivamente a las categorías respiratory- disease_651_4192_n y cancer_651_4195_n. Ambas están incluidas en el *grupo* disease_651_n.

Según esta tipografía, en el fragmento de cadena que comentamos hay seis relaciones de identidad (ID), cuatro de grupo (GRP) y tres categoriales (CAT). Según Ellman (1998), el 60-80% de las cadenas léxicas se deben a la relación ID, que se basa en la repetición léxica.

8.1.4. Representación de cadenas léxicas *(Chains)*

El primer lexema de cada cadena nos remite a una página HTML que representa las cadenas léxicas individuales. Esta *representación* especifica las palabras en esa cadena, el tipo de vínculo léxico entre ellas (ID, CAT, GRP, ONE) y las categorías del tesauro mediante las que se establecen los vínculos. Si nos situamos con el ratón encima de lung-cancer[1], que inicia la cadena 1, una de las 14 cadenas del texto, encontraríamos 276 lexemas correspondientes a 23 formas léxicas distintas (véase Apéndice VIId). Presentamos una sección de esta cadena:

TABLA 32: Fragmento de una de las 14 cadenas léxicas del texto QDT2.

Chain 1

lung-cancer, lung-cancer, lung-cancer, carcinoma, carcinoma, symptoms, patients, disease, patients, diagnosis, patients, disease, patients, prognosis, patients, disease, patients, lung-cancer, patients, disease, patients, diagnosis, symptoms, patients, patients, confined, symptoms, patients, prognosis, symptoms, patients, lung-cancer, patients, patients, patients, patient, lung-cancer, lung-cancer, critical, cases, lung-cancer, carcinoma, cancers, cancers, carcinoma, cancer, lesion, cancer, patients, cancer, patients, lung-

cancer, lung-cancer, carcinoma, carcinoma, carcinoma, carcinoma, critical, patients, critical, disease, patients, patients, patients, lung-cancer, lung-cancer, cancer, cancer, patient, lesion, lesions, lesions, lesion, cancer, bronchial, carcinoma, bronchial, lung-cancer, patients, cases, patient, carcinoma, lung-cancer, cancers, patients, disease, patients, disease, disease, cancers, lung-cancer, lung-cancer, lung-cancer, patient, lung-cancer, patients, disease, lung-cancer, lung-cancer, carcinoma, cancers, patients, patients, cancers, lung-cancer, patients, lung-cancer, patient, condition, patient, critical, patients, lung-cancer, patients, cancer, patients, patients, patients, patients, patients, inoperable, patients, disease, patients, inoperable, patients, inoperable, patients, patients, critical, patients, patients, smokers, lung-cancer, patients, cancers, cancers, cancers, cancers, patients, cancers, cancers, patients, patients, patients, lung-cancer, patients, lung-cancer, patient, condition, patient, critical, inoperable, patients, disease, patients, patients, inoperable, patients, disease, disease, critical, patients, patients, patients, carcinoma, patients, disease, patients, lung-cancer, patients, lung-cancer, patients, patients, patients, patients, disease, patient, patients, disease, patients, patients, patients, patients, patients, cancer, patients, patients, patients, lung-cancer, patients, patients, disease, cases, disease, cases, patients, patients, cases, patients, lung-cancer, patients, lung-cancer, patients, patients, patients, disease, patients, patient, patients, disease, cough, chest-pain, patient, patients, patients, disease, patients, inoperable, patients, patient, patients, disease, patients, lung-cancer, toxic, patients, patients, patients, lung-cancer, lung-cancer, patients, inoperable, disease, lung-cancer, patients, patients, lesions, bronchial, pain, cases, lesions, critical, patient, patient, lesion, toxic, patients, lesions, lung-cancer, patients, lung-cancer, patients, lesion, disease, toxic, patients, patients, patients, patients, patients, patients, carcinoma, patients, cancers, lesion, cancer, patients, lesion, patients, patients, disease, patients, patients, lesions, patients,

:lung-cancer :Wd 4, Sent 1, Para 1, ???-> 0, Value 0
respiratory-disease_651_4192_n
:lung-cancer :Wd 11, Sent 1, Para 3, ID-> 4, Value 0
respiratory-disease_651_4192_n
:lung-cancer :Wd 25, Sent 1, Para 3, ID-> 11, Value 0
respiratory-disease_651_4192_n
:carcinoma :Wd 31, Sent 1, Para 3, GRP-> 25, Value 0
cancer_651_4195_n
:carcinoma :Wd 36, Sent 1, Para 3, ID-> 31, Value 0
cancer_651_4195_n
:symptoms :Wd 65, Sent 1, Para 3, GRP-> 36, Value 0
illness_651_4186_n
:patients :Wd 81, Sent 1, Para 3, GRP-> 65, Value 0
sick-person_651_4202_n
:disease :Wd 84, Sent 1, Para 3, GRP-> 81, Value 0
disease_651_4187_n
:patients :Wd 95, Sent 1, Para 4, ID-> 81, Value 0
sick-person_651_4202_n
:diagnosis :Wd 97, Sent 1, Para 4, GRP-> 95, Value 0
pathology_651_4203_n
:patients :Wd 98, Sent 1, Para 4, ID-> 95, Value 0
sick-person_651_4202_n
:disease :Wd 112, Sent 1, Para 4, ID-> 84, Value 0
disease_651_4187_n
:patients :Wd 120, Sent 1, Para 4, ID-> 98, Value 0
sick-person_651_4202_n
:prognosis :Wd 139, Sent 1, Para 4, CAT-> 97, Value 0
pathology_651_4203_n

Se especifica con qué lexema previo está vinculado cada elemento de la cadena, que es el precedido por la flecha (->). Por ejemplo, la undécima palabra del texto (Wd 11), *lung-cancer*, está vinculada mediante la relación de identidad con la cuarta palabra, *lung cancer* (ID-> 4). Las referencias al número de oración y de párrafo no son exactas, un aspecto que el autor del programa piensa modificar[107]. Después de esto, encontramos la entrada del tesauro bajo la que aparece, acompañada del número que se le atribuye en el tesauro y de una letra (n) que hace mención a la categoría gramatical: respiratory-disease_651_4192_n.

Por último, aclararemos que las cadenas identificadas representan el 70 % de los conceptos presentes en el *perfil.* Es decir, además de estas cadenas, habría otras que o bien son menos importantes o bien se establecen entre lexemas separados por más de 500 palabras. Estas contribuyen al restante 30 % de la cohesión del texto.

8.2. Limitaciones de los resultados aportados por *Hesperus*

El *output* de *Hesperus* no es suficiente para plasmar las relaciones cohesivas del texto. Como acabamos de ver, no se explicitan todas las cadenas y no se computan las relaciones entre lexemas separados por más de 500 palabras. Además, a diferencia del lector humano, un ordenador no detecta la interacción entre la maquetación del texto, por una parte, y el género textual, la deixis, la estructura discursiva y la estructura informativa (*tema - rema; información principal - información secundaria*) del texto, por otra. Esto se constata cuando el ordenador pasa por alto que las palabras en negrita son más importantes desde el punto de vista cohesivo que las que no lo están. Si esta interacción ayuda a inferir más rápidamente las palabras clave sobre las que gira la cohesión, podemos decir que el ordenador carece de esta fuente de información sobre la cohesión del texto.

En relación con el tesauro, nos gustaría señalar que, por ser de la lengua general, no contiene todos los términos propios de la oncología, muchos de los cuales son UF que *Hesperus*, salvo en contadas ocasiones, no tiene la capacidad de reconocer. Y así,

[107] Comunicación personal.

vemos cómo se identifica la UF *lung cancer* pero no otras muy frecuentes en oncología como, por ejemplo, *small cell lung cancer* o *magnetic resonance imaging*.

Por otra parte, los fenómenos de polisemia, homonimia, ambigüedad y connotación presentes en el lenguaje dificultan la búsqueda de casos de cohesión por parte de un ordenador. Aunque en el lenguaje científico estos fenómenos se producen con menos frecuencia, seguimos con la limitación de utilizar un glosario de la lengua general donde los términos no están estructurados de acuerdo con el conocimiento sobre oncología. En muchas ocasiones *Hesperus* no capta el contexto y el significado que adquieren estos términos en los textos sobre cáncer. Por ejemplo, en la cadena 13 del texto QDT1, *nodes* queda asociado al concepto COMPOSICIÓN, que es la idea imprecisa que subyace a las distintas acepciones del término *node*, perteneciente a diferentes campos del saber: botánica, anatomía, astronomía, física, etc. En textos sobre oncología, designa unos componentes del sistema linfático, los ganglios linfáticos, a los que emigran con bastanta facilidad las células cancerosas de un tumor cercano.

> **node n.**
> **1** Bot. a the part of a plant stem from which one or more leaves emerge. b a knob on a root or branch.
> **2** Anat. a natural swelling or bulge in an organ or part of the body.
> **3** Astron. either of two points at which a planet's orbit intersects the plane of the ecliptic or the celestial equator.
> **4** Physics a point of minimum disturbance in a standing wave system.
> **5** Electr. a point of zero current or voltage.
> **6** Math. a a point at which a curve intersects itself. b a vertex in a graph.
> **7** a component in a computer network.
>
> *(Concise Oxford Dictionary)*

Tampoco percibe los solapamientos entre cadenas que se derivan de la realidad extralingüística, que en el ámbito de la ciencia está en continuo cambio. *Hesperus* no señala unidades léxicas susceptibles de aparecer en dos cadenas diferentes. Sin embargo, este fenómeno de doble pertenencia es frecuente dado que las palabras no tienen límites perfectamente definidos. De hecho, cuando para cada texto hagamos nuestra propuesta definitiva de cadenas léxicas, marcaremos con vínculos de hipertexto aquellas palabras que pertenecen a más de una cadena. Por ejemplo, el término *clinical trials* pertenece al mismo tiempo a las cadenas léxicas *TREATMENT* y *RESEARCH*.

Para que un programa informático detecte la cohesión en un texto especializado de un modo eficaz, sería preciso adaptar su tesauro al subdominio al que pertenece y actualizarlo cada cierto tiempo. Los diferentes sentidos de los lexemas de la lengua general quedarían restringidos a los sentidos que adquieren en ese campo del saber. Además, en el caso de lexemas polisémicos, el concepto asociado se determinaría en función de las palabras de su radio colocacional. Esta tarea de desambiguación, que nos hará modificar las cadenas léxicas, requiere la ayuda de un programa de análisis léxico que elabore concordancias y listas de frecuencia. En estas últimas, queda patente la repetición de la misma forma léxica, por lo cual, reflejan muy bien el papel de la repetición en la creación de la cohesión. Se empleará para este fin el programa *Wordsmith Tools.*

Para ilustrar estas limitaciones, mencionaremos un ejemplo tomado de una de las cadenas reconocidas en el texto QDT2, en la que predomina el concepto MINUTENESS. En ella se asocian lexemas que, si bien comparten un sema que indica pequeñez, en el campo de la oncología, su significado básico no gira en torno de este sema. Este es el caso de los lexemas en cursiva, cuyo significado básico queda modificado por su cotexto, de forma que se encuadrarán en las áreas entre paréntesis:

Fractionation (RADIOTHERAPY)
Bronchial *compression* (SYMPTOMS)
Microscopic examination (DIAGNOSIS)

En cuanto al formato de presentación de los resultados de *Hesperus*, opinamos que su exhaustividad plantea una dificultad a la hora de plasmarlos en soporte papel. Por ello, los adaptaremos a nuevos formatos en 8.3.

Por todo esto concluimos que, si bien el programa ha sido de gran utilidad, hemos encontrado algunos puntos débiles, que intentaremos subsanar con una presentación de los resultados más informativa y visual para el lector y modificando las cadenas léxicas identificadas de acuerdo con las áreas conceptuales de la oncología y con los datos aportados por *Wordsmith Tools.*

8.3. Etapas para el análisis extensivo de la cohesión

Teniendo en cuenta las limitaciones expuestas, proponemos un análisis de la cohesión que destaque los factores que intervienen en la creación de la misma: la textualidad, las relaciones conceptuales y la repetición. A partir de estas, sugerimos cinco etapas que aplicaremos al análisis cohesivo de seis textos sobre el tratamiento del cáncer de pulmón (véase capítulo 11).

a. Índice del texto
b. Áreas conceptuales reconocidas a partir de listas de frecuencia lematizadas
c. Representación modificada de los resultados de *Hesperus*
d. Propuesta definitiva de cadenas léxicas
e. Cuadro contrastivo de las diferencias entre la cohesión detectada por un ordenador y por un humano

Los resultados de *Hesperus* se volcarán en la tercera etapa, aunque adaptados al formato que hemos propuesto en aras de la claridad y la concisión. La cuarta etapa integra los resultados de las que la preceden.

8.3.1. Índice del texto

Todos los textos analizados comienzan con un índice con hipervínculos hacia las diferentes secciones en que se divide cada uno de ellos. Como hemos visto, existe una relación entre la cohesión, por una parte, y la estructura y segmentación de textos, por otra (Berber Sardinha 1997). Por eso, sospechamos que las secciones que aparecen en el índice de un texto apuntan hacia las grandes áreas conceptuales explotadas en el mismo y materializadas en cadenas léxicas. En definitiva, al observar el índice, nos familiarizamos con la faceta textual de la cohesión.

8.3.2. Áreas conceptuales reconocidas a partir de listas de frecuencia lematizadas

Cuando se somete una lista de frecuencia creada automáticamente a lematización se ve un claro predominio de determinados lexemas. Intuimos que estos son los que sirven de hitos en el campo conceptual de la oncología y en sus posibles marcos. En

consecuencia, para cada texto elaboraremos una lista con las áreas conceptuales bajo las que se agrupan las unidades léxicas y terminológicas con una frecuencia relativa mayor o igual al 0,1%. Dado que no incluye todas las palabras y términos del texto, esta lista será provisional. Para cada texto, la modificaremos en las secciones *Propuesta definitiva de cadenas léxicas* del capítulo 11.

Para identificar las grandes áreas conceptuales de cada texto, se exportarán los datos de las listas de frecuencia creadas con *Wordsmith Tools* a la hoja de cálculo *Microsoft Excel*, de forma que se ubiquen en la misma columna los lexemas relacionados semánticamente. Por ejemplo, en el texto QDT2, los lemas que denominan alguna parte del cuerpo tanto a nivel anatómico como histológico, aparecerían en la misma columna. Su porcentaje con respecto a las demás áreas conceptuales indica de un modo preliminar la importancia de los conceptos relacionados con alguna PARTE DEL CUERPO.

TABLA 33: Activación de términos que designan partes del cuerpo con frecuencia mayor al 0,1% en QDT2 (*Microsoft Excel)*.

LEMA	%
LUNG	1,12
CELL	0,79
CHEST	0,25
BRONCHUS	0,23
BRAIN	0,17
LYMPH NODES	0,16
PLEURA	0,14
LOBE	0,11
TOTAL	2,97

La suma de todos los porcentajes de la columna nos indicaría aproximadamente la representación del área conceptual PARTS OF THE BODY en ese texto (2,97%). Un diagrama de sectores con los porcentajes atribuidos inicialmente a cada área conceptual ilustrará los marcos conceptuales preliminares del texto (Véase, por ejemplo, *11.1.2.2. Áreas conceptuales reconocidas a partir de listas de frecuencia lematizadas*).

8.3.3. Representación modificada de los resultados de *Hesperus*

Se aprehenderán buena parte de las relaciones conceptuales que contribuyen a la cohesión con la ayuda de la *perspectiva general*, el *perfil* y la *representación de las cadenas léxicas* generadas por *Hesperus.*

En primer lugar, a partir de la *perspectiva general* y el *perfil*, describiremos cuantitativamente la *calidad cohesiva* del texto. Sabremos cuántas palabras plenas hay en el texto y cuántas categorías del tesauro *Roget*, y atribuiremos al texto un coeficiente de cohesión[108] dividiendo la puntuación bruta (*raw score*) entre el número total de palabras del texto. Asimismo, especificaremos las categorías del tesauro que representan más del 1% de la cohesión del texto. Muchas de estas categorías habrán ya sido identificadas en la segunda etapa, *Áreas conceptuales reconocidas a partir de listas de frecuencia lematizadas*. Estas categorías son el nexo de unión entre los distintos lexemas que constituyen las cadenas identificadas por *Hesperus.*

Seguidamente, analizaremos cada una de estas cadenas con el fin de confeccionar tablas donde se indica la frecuencia absoluta de las formas léxicas vinculadas. Las tablas tienen la siguiente presentación (véase Apéndice VII e)[109]:

TABLA 34: Representación tabular de las cadenas reconocidas por *Hesperus* (Cadena 4 del texto QDT2)

NODES	7
NODE	4
NODULES	3
BONE	1

CHEST	1
NODULE	1
OSSEOUS	1

[108] Como se ha explicado con anterioridad, este valor es dependiente de la extensión del texto, por lo que para no distorsionar la información que proporciona sobre la fuerza de las cadenas y su extensión, siguiendo el consejo del profesor Ellman, lo hemos dividido entre el número total de palabras.

[109] La unidad léxica en negrita es la que inicia la cadena.

Asignaremos a cada cadena una etiqueta que refleja el contenido semántico que comparten los lexemas que la componen. Para ello, se examinarán las categorías conceptuales de cada cadena (por ejemplo, respiratory-disease_651_4192_n) y las listas de frecuencia lematizadas. Con estos datos y guiados por nuestra intuición, hemos etiquetado la mayoría de las cadenas a excepción de algunas que incluyen palabras relacionadas semánticamente de una forma indirecta. En estas cadenas dudosas, las relaciones de *identidad* son menos frecuentes que las de *categoría* o las de *grupo*, al contrario de lo que ocurre en las cadenas que hemos podido etiquetar, en las que predomina la relación de *identidad* o *categoría*. Además, vinculan normalmente lexemas que pertenecen al metadiscurso científico. A esas cadenas las denominaremos *cadenas difusas*, las etiquetaremos con el signo '?' y para cada uno de sus elementos, explicitaremos la categoría del tesauro que justifica su inclusión en esa cadena. Intentaremos subsanar las inevitables incoherencias y las agrupaciones erróneas de lexemas derivadas de la ambigüedad y la polisemia.

Comentaremos estas cadenas, aunque nos centraremos, más que en los lexemas, en los conceptos activados y su porcentaje de activación. Para representarlos, se elaborarán tablas similares a esta:

TABLA 35: Representación de los conceptos identificados por *Hesperus* en las cadenas 4, 7 y 11 de QDT2: porcentaje de activación.

CONCEPTOS ACTIVADOS					
Cadena 4 *(Nodules)*		**Cadena 7** *(Lymph Nodes)*		**Cadena 11** *(Pleura)*	
SWELLING	1,56	**BLOOD**	0,68	**LATERAL** (Pleura)	0,76
HARD	0,13	**SANGUINEOUS**	0,22	**LATERALITY**	< 0,1
BOSOM	< 0,1	**FLUID**	0,16		
HARDNESS	< 0,1				
TOTAL	+1,69	**TOTAL**	1,06	**TOTAL**	+0,76
↓		↓		↓	
HARD PARTS OF THE BODY		**BODY FLUIDS**		**PLEURA**	

Como se puede apreciar, hemos agrupado estas cadenas en bloques de dos o tres en aquellas ocasiones en las que exista una relación entre ellas. En esta representación junto

al número de cadena concedido por *Hesperus* (*Cadena 4*), está en cursiva el lexema que inicia la cadena (*nodules*). Abajo, están las categorías del tesauro y su porcentaje de activación. En la casilla que sigue la flecha, está la etiqueta que hemos atribuido a cada cadena. Según esta tabla, habría en el texto tres cadenas en torno al concepto BODY, al que *Hesperus* atribuye un 3,51%. Este porcentaje no dista demasiado del que se infería de la lista de frecuencia lematizada (2,97%) del apartado 8.3.2.

8.3.4. Propuesta definitiva de cadenas léxicas

Con toda la información recabada en los apartados anteriores y teniendo en cuenta la forma en que se estructuran los términos en el campo de la oncología, creemos conveniente hacer una formulación definitiva de las cadenas léxicas presentes en cada texto. Para llegar a esta propuesta, se eliminarán los lexemas incluidos erróneamente, se ubicarán en cadenas nuevas, se añadirán unidades léxicas nuevas a las cadenas ya existentes o fusionaremos cadenas y crearemos nuevas cadenas y subcadenas. Asimismo, prestaremos una especial atención a las unidades lexicalizadas y fraseológicas que aparecen.

En las cadenas, no especificaremos la forma canónica de las unidades léxicas ya que nuestra aproximación a la terminología es descriptiva y nos interesa el uso de los términos en nuestro corpus y las distintas variaciones que se producen con respecto al término más frecuente. Es decir, nuestro interés está en el uso contextual de los términos y no en su uso prescriptivo y estandarizado. En consecuencia, en muchas ocasiones los verbos estarán en la forma *'-ed'*, en lugar de en infinitivo, y los sustantivos en plural. Esto nos indica que la forma canónica no tiene por qué ser la más frecuente.

Por otra parte, bajo la misma entrada, a veces introduciremos verbos, sustantivos, adjetivos o adverbios. Esto es así porque la cohesión se crea independientemente de la categoría gramatical de las unidades léxicas vinculadas y depende principalmente de la repetición conceptual y léxica. Y así, uno de los componentes de la cadena LOCATION IN THE HUMAN BODY del texto QDT2 es el lema *regional* bajo el que se agrupan las formas con las que establece cohesión mediante repetición exacta (*regional-regional*) y repetición variada sintáctica (*regional-region; regional-regionally*). Se especifican las UF que forma

en el texto, que se relacionan mediante repetición sintagmática exacta *(regional lymph node-regional lymph node)* y simple *(regional lymph node-regional lymph nodes).*

TABLA 36: Casilla perteneciente a una de las *cadenas léxicas definitivas* de QDT2

REGIONAL Regional 9 ~ lymph node(s) 3 Region 1, regionally 1	11

La creación de nuevas cadenas obedece a que tanto las listas lematizadas como el perfil del documento apuntan hacia áreas conceptuales significativas en el texto y en el campo de la oncología, pero que *Hesperus* no ha reconocido. Hemos seguido una serie de criterios que justifiquen la creación, agrupación o eliminación de cadenas:

1. Se creará una cadena siempre *Hesperus* la haya creado en otro texto relacionado del corpus, con tal de que el porcentaje asignado al concepto por el perfil del documento lo respalde.
2. Se tendrá en cuenta las áreas conceptuales activadas en el evento médico oncológico (Faber 1997 y Tercedor 1999).
3. Son candidatos a formar cadena tanto los diez primeros conceptos en el perfil del documento como los diez primeros lexemas de las listas de frecuencia que no se puedan integrar dentro de otros.

8.3.5. Cuadro contrastivo (*Hesperus* – Análisis combinado)

Como última etapa, se apreciarán las diferencias entre la cohesión susceptible de ser detectada por un ordenador con tesauro y la detectada por un lector. La primera es principalmente explícita (repetición léxica) y asocia lexemas de la lengua general. En cuanto a la segunda, se basa no sólo en la repetición léxica sino también en el conocimiento especializado, que comprende conocimientos de terminología médica, del método científico y de los principales recursos textuales en ese subdominio. Se presentarán estas diferencias cuantitativamente en una tabla que compara las cadenas identificadas por *Hesperus* y nuestra propuesta. Esta tabla incluye el número de lexemas

vinculados, el número de formas distintas y la activación conceptual correspondiente a cada cadena reconocida por *Hesperus* (columna izquierda) o por nosotros (columna derecha). Esta información va a veces entre corchetes.

TABLA 37: Fragmento del cuadro contrastivo correspondiente al texto QDT2.

CADENAS LÉXICAS RECONOCIDAS POR *HESPERUS*		**NUESTRA PROPUESTA DE CADENAS LÉXICAS**	
H 3: TREATMENT, DIAGNOSIS, MEDICINE [105-23-21,39] **H 1: DRUGS** [6-2-4]			**0. TREATMENT** [143-30 -23,6] **4. DIAGNOSIS** [53-24-8,74]
Lexemas vinculados	111	196	Lexemas vinculados
Formas distintas	25	54	Formas distintas
Activación conceptual ajustada (%)	29,52	32,34	Activación conceptual (%)
ERA (0,65)			**11. TIME** [8-8-1,32]
Activación conceptual ajustada (%)	0,76	1,32	Activación conceptual (%)
[...]			**[...]**
TOTAL			
		606	Suma de elementos en las cadenas
Lexemas vinculados	345	552	Lexemas vinculados[110]
Formas distintas	80	192	Formas distintas[111]
Activación conceptual según Hesperus	86,01		
Activación conceptual ajustada (%)	100	99,96	Activación conceptual (%)
DATOS ADICIONALES SOBRE EL TEXTO QPT3 Nº total de palabras: 1255 Palabras que crean cohesión (*Hesperus*): 562		Categorías conceptuales del tesauro: 83 Coeficiente de cohesión (*Hesperus*): 12,21 Nº de oraciones: 66	

Ha sido preciso ajustar los porcentajes conceptuales de *Hesperus* porque la suma de estos no llegaba al 100%. También se han computado los porcentajes de conceptos que no forman cadena pero que aparecían en el perfil del documento. Este es el caso del concepto ERA, que no forma cadena pero que tiene una activación conceptual (0,76%) no muy distinta a la que le hemos dado nosotros (1,32%).

[110] Sólo se ha computado una vez los lexemas que aparecen en más de una cadena.
[111] De estas, hay 21 formas léxicas que pertenecen a más de una cadena y que llevan asociadas 54 lexemas adicionales a los 552 lexemas vinculados.

Los porcentajes que aportamos como resultado de nuestro análisis se calculan teniendo en cuenta el número de lexemas de cada cadena léxica (143 en la cadena *TREATMENT*) en relación con la suma de los elementos de todas las cadenas (606). Esta suma es superior al número de lexemas vinculados (552) ya que, como veremos, hay lexemas susceptibles de ser clasificados en varias cadenas. La diferencia entre las dos indica el número de lexemas que hacen de nodos de más de una cadena.

Por último, se añaden datos estadísticos obtenidos mediante *Wordsmith Tools* (número total de palabras, nº de oraciones) y *Hesperus* (palabras que crean cohesión, categorías conceptuales del tesauro *Roget* y coeficiente de cohesión). El número de palabras que crean cohesión según *Hesperus* (562) no es muy distinto al que proponemos en la casilla de la derecha, *Lexemas vinculados* (552).