11.3. Datos estadísticos relacionados con el estudio de la cohesión y sus limitaciones

En esta sección se comentan las estadísticas correspondientes a los textos para profesionales de la salud y para pacientes, comentadas en 11.1. y 11.2., para ver si la cohesión se construye de forma distinta en distintos tipos textuales definidos con respecto al parámetro "grado de conocimiento del lector". También se establece la relación entre el formato con el que hemos presentado las cadenas léxicas y los tipos de repetición propuestos en el capítulo 6.

11.3.1. Representación de las cadenas léxicas y tipos de repetición

El formato de las tablas incluidas en las secciones tituladas *Desarrollo de las cadenas léxicas definitivas* de este capítulo son indicativas del tipo de relación cohesiva establecida, de acuerdo con una serie de consideraciones. Las aplicamos a la subcadena *GENERIC TREATMENT* de QDT1 (tabla 120):

- Para detectar las instancias de *repetición exacta* en una cadena/subcadena, se le resta al TOTAL el número de formas que tienen una frecuencia absoluta de 1. En la subcadena *GENERIC TREATMENT* encontraríamos 40 instancias de repetición exacta, resultantes de restar a 45, las 5 variantes morfológicas que sólo aparecen una vez (*treatments, curative, modality, option, procedures*). Para todo el texto, a partir de una lista de frecuencia en la que se han suprimido las palabras forma, sólo basta con restar al número de ocurrencias (*tokens*), el número de formas que tienen una frecuencia igual a 1.
- El encontrar formas distintas (*types*) en una casilla indica que entre ellas existe *repetición simple* y *variada sintáctica*. Las formas *treatment* y *treatments* están relacionadas mediante repetición simple y estas lo están a su vez con *treated* mediante repetición variada sintáctica. Para computar los casos de repetición simple y variada sintáctica en una subcadena, se resta al número de formas distintas (14) el número de lemas (9). Por tanto, en esta cadena habría 5 casos entre repetición simple (*treatment – treatments; modalities – modality; options – option*) y variada sintáctica (*treatment – treated; cure – curative*).
- El número de lemas de una cadena/subcadena queda reflejado en el número de casillas. El número de lemas indica el mínimo de instancias de *repetición*

paradigmática semántica, dado que todos esos lemas están relacionados mediante el concepto que etiqueta la cadena. En la subcadena de la tabla 120 hay nueve lemas y relaciones paradigmáticas sinonímicas (TREATMENT – THERAPY, TRIALS – TREATMENT), hiponímicas (APPROACH – THERAPY, PROCEDURES – THERAPY) y asociativas circunstanciales basadas en la relación causa-efecto (TREATMENT– CURE, TREATMENT– CONTROL).

- Cada vez que aparece una unidad poliléxica o los signos ~ o ➙, seguidos de un número igual o mayor que 2 nos encontramos ante casos de *repetición sintagmática* o *repetición mixta*. En la tabla 120, podemos señalar *surgical approaches, treatment approaches, clinical trials, treatment options, intracavitary treatment.*

TABLA 120: Subcadena *TREATMENT* del texto QDT1.

0. TREATMENT (H1, H3a, H10, H13)

0.1. Generic treatment

TREATMENT Treatment 12, treated 5, treatments 1	18
➙ APPROACHES Surgical ~ 2 Treatment ~ 2 Physical ~ 1	5
TRIALS (See 1.1.) Clinical ~ 4	5
CONTROL (+ SYMPTOMS+)	4
CURE (See 4.2.) Cure 2, curative 1	3
MODALITIES (See 11.) Modalities 2, combined modality 1	3
OPTIONS (See 11.) Treatment options 2 Option 1	3
THERAPY Intracavitary ~ 2 (See 0.4.3. RADIATION ~)	3
➙ PROCEDURES	1
TOTAL	45

11.3.2. Estadísticas de textos sobre tratamiento del cáncer de pulmón

Si en los textos sobre tratamiento para profesionales, QDT1, QDT2 y QDT3, el porcentaje medio de palabras que contribuyen a la cohesión del texto (52,64 - 46,6 - 50,85) es de 50,03%, en los textos para pacientes, QPT1 (46,21%), QPT2 (44,67%) y QPT3 (43,98%), este es ligeramente menor, en concreto, 44,95%. Esto se debe principalmente a dos factores.

Por una parte, en inglés, la nominalización típica de textos científicos para especialistas toma cuerpo en *sintagmas nominales compuestos* (Bhatia 1993: 149). La estructura de estos sintagmas nominales es la de *modificador(es)* + *núcleo*. La posición que precede al sustantivo nuclear queda ocupada por una sucesión lineal de sustantivos, entre los que se puede intercalar algún adjetivo. Veamos algunos ejemplos sacados de los textos QDT2 y QDT3:

(98) Non-small cell lung cancer (QDT2)
(99) Squamous cell lung cancer (QDT2)
(100) Ipsilateral hilar lymph nodes (QDT2)
(101) Endobronchial laser therapy (QDT3)
(102) Magnetic resonance imaging scans (QDT3)

Esta estructura es muy frecuente en las unidades fraseológicas de los textos para profesionales de la salud y hace que haya una mayor profusión de palabras plenas, y por tanto, un porcentaje mayor de palabras que contribuyen a la cohesión del texto, ya que los sintagmas nominales se construyen sin la necesidad de preposiciones o pronombres relativos.

Por otra parte, en los textos para especialistas, nuestros cálculos indican que la repetición léxica exacta y simple es más habitual que en los textos para pacientes. Como estos recursos son totalmente explícitos, se han computado todos los casos, con lo que se obtiene un porcentaje elevado de palabras cohesionadas, algo que no podemos afirmar en el caso de la repetición variada léxica, más frecuente en textos para pacientes pero más difícil de detectar.

En los textos para especialistas se observa una mayor repetición exacta, destinada a evitar la ambigüedad, mientras que existe una mayor variedad léxica en los textos para pacientes, de acuerdo con los cálculos que hicimos en torno al número de variantes morfológicas y de lemas cada 100 palabras. Se dejará a un lado el texto QDT1, en el que se produce una variedad léxica inusual debido a su estilo esquemático, sobre todo en el apartado *Treatment options*.

En QDT2 y QDT3, por cada 100 palabras, hay respectivamente 24 y 27 formas léxicas distintas correspondientes a 16 y 19 lemas distintos. La variedad léxica es mayor en la versión para pacientes, en concreto, por cada 100 palabras hay una media de 35,3 formas léxicas distintas asociadas a 31,3 lemas. Esto se explica por el fenómeno que cita Meyer (1991: 5) por el que en textos de contenido científico destinados a un lector lego, en este caso a pacientes, la cohesión se logra mediante sinónimos y términos subordinados y superordinados, y no tanto mediante la repetición de la misma forma léxica. Profundizaremos más sobre esta idea en el capítulo 13.

Otra explicación a la variedad léxica y al elevado porcentaje de unidades léxicas que contribuyen a la cohesión la hallamos en la utilización de paréntesis en la versión para pacientes, un recurso muy sinténtico con el que se activan un gran número de conceptos con el mínimo de palabras, como se verá también en el capítulo 13.1.

(103) The prognosis (chance of recovery) [QPT2]

(104) The area that separates the two lungs (mediastinum) [QPT2]

Con el paréntesis se consigue un aumento de la *type-token ratio*, y por eso, este índice es anormalmente elevado en los textos para pacientes.

A partir del cómputo que hace *Hesperus* no sólo de las categorías del *Roget* activadas sino también del número de palabras que crean cohesión en los textos analizados, se puede hallar un promedio de las categorías conceptuales relacionadas con cada 100 palabras vinculadas. Este promedio no difiere apenas del calculado en 11.1. y 11.2. a partir de nuestro análisis:

Categ. conceptuales por cada 100 palabras activadas

QDT1 - QDT2 - QDT3:	QPT1 - QPT2 - QPT3:
25 – 8 – 11 (*11.1. y 11.2.*)	14 – 14 – 15 (*11.1. y 11.2.*)
28 – 8 – 11 (*Hesperus*)	15 – 14 – 15 (*Hesperus*)

La obstención de este cálculo mediante métodos informáticos permite el estudio de la densidad conceptual correspondiente a un mismo número de palabras plenas. Además, permite su estudio no sólo en seis textos sobre el mismo tema que se diferencian respecto al parámetro *destinatario* sino en un corpus de 36 textos. Como se verá en 11.3.3., en los textos destinados a un lector lego, se necesitan menos palabras plenas para activar las principales áreas conceptuales que aparecen en los textos para especialistas. Estas palabras plenas son en su mayor parte unidades léxicas de la lengua general que "arropan" aquellos términos que pueden crear dificultad al lector.

11.3.3. Estadísticas generales en el resto de los textos

Si se extienden al resto del corpus del *Physician Data Query* algunos de los cálculos aplicados a los textos sobre tratamiento, es posible identificar tendencias estadísticas que puedan confirmar diferencias cohesivas entre textos para especialistas y textos para lectores legos. Claro está, aparecen las limitaciones que impone la composición del corpus analizado. La tabla de datos estadísticos incluye los siguientes parámetros:

a) *type / token ratio (x100)*: indica el número de formas distintas por cada 100 palabras de texto, es decir, el porcentaje entre variantes morfológicas y total de palabras. Su valor depende en gran medida de la extensión del texto, de forma que, cuanto mayor sea su extensión, su *type / token ratio* es menor. Para que no quede distorsionado, en la casilla correspondiente habrá dos cálculos. El primero se basará en una lista de frecuencia sin modificar y el segundo, en una lista en la que se han eliminado las palabras forma con la ayuda de una *stoplist* que contiene 256 palabras. Como resultado de esto, en el corpus para especialistas se han eliminado 130 palabras forma (1,88% del total), mientras que en el destinado a pacientes se ha prescindido de 208 palabras (6,14% del total). Asimismo, para todo el corpus se ha calculado la *standardised*

type/token ratio[169] por cada 300 palabras de texto, puesto que el corpus cuenta con algunos textos que superan escasamente las 600 palabras.

b) *Palabras vinculadas por cada 100 palabras de texto*: para este cálculo, nos serviremos del primer valor que otorga *Hesperus* en la perspectiva general (p. 707).
c) *Coeficiente de cohesión*: es el cociente que se obtiene al dividir entre el total de palabras, la cifra que *Hesperus* otorga a cada texto en función del número de conceptos del *Roget* activados (segunda cifra entre paréntesis del Apéndice VIIa). Como esta cifra depende de la extensión del texto, estimamos que la fiabilidad de este cálculo es escasa.
d) *Palabras vinculadas en cada oración*: dividiremos el número de palabras vinculadas entre el número de oraciones para ver cuántas palabras hay vinculadas por oración. Se relacionarán estos cálculos con la extensión de las oraciones, que es mayor en textos para especialistas (22,15 palabras) que en textos para pacientes (17,71).
e) *Categorías conceptuales del tesauro Roget por cada 100 palabras vinculadas:* con este cálculo se pretende detectar qué textos son más informativos en relación con el número de conceptos que activan y las palabras plenas que los activan.
f) *Número de cadenas*: no se prestará atención a este cálculo porque existe una gran disparidad entre los textos. La mayor extensión de los textos QD hace que *Hesperus* reconozca en ellos más cadenas, una media de 26,5 cadenas, en comparación con el promedio de 7,83 cadenas que reconoce en QP.

El coeficiente de cohesión y el número de categorías conceptuales por cada 100 palabras vinculadas es sólo orientativo. Somos conscientes de las limitaciones de estos cálculos ya que el tesauro es de la lengua general y no permite la identificación de unidades léxicas exclusivas de la medicina, y por tanto, excluye relaciones conceptuales que se dan en el texto.

[169] Este cálculo es la media de las *type/token ratios* que se obtienen a partir de *n* palabras consecutivas de texto. Si se fija *n* con el valor de 300, se estima la *type/token ratio* de las primeras 300 palabras, después de las 300 siguientes, y así sucesivamente, y luego se hace una media que es a lo que se le denomina la *standardised type/token ratio*.

TABLA 121: Estadísticas correspondientes al corpus analizado en el capítulo 11.

	PROFESIONALES DE LA SALUD (QD)					**PACIENTES (QP)**				
	Type/ token ratio(x100) (con / sin pal. forma)	**Palabras vinculadas (x100) / n. palabras**	**Coeficiente de cohesion**	**Palabras vinculadas/ n. de oraciones**	**Cat. conceptuales por 100 pal. vinculadas**	**Cat. conceptuales por 100 pal. vinculadas**	**Palabras vinculadas/ n. oraciones**	**Coeficiente de cohesion**	**Palabras vinculadas (x100) / n. palabras**	**Type/ token ratio(x100) (con/ sin pal. forma)**
OVERALL	Normal 9,68 9,5	46,55	9,73	10,71	13,41	17,26	8,69	11,93	45,86	Normal 9,50 8,92
	Estánd 56,04									Estánd 50,22
Malignant mesothelioma (T1)	38,92 32,44	505/ 1079= 46,8	9,76	505/ 63= 8,01	141/ 505= 27,92	109/ 704= 15,48	704/ 92= 7,65	10,7	704 / 1621= 43,42	24,80 19,31
Non-small cell lung cancer (T2)	18,23 16,15	2470/ 5523= 44,72	8,37	2470/ 251= 9,84	196/ 2470= 7,93	80/ 553= 14,46	553/ 85= 6,50	13,83	553 / 1258= 43,95	23,37 16,93
Small cell lung cancer (T3)	21,27 18,37	1648/ 3408= 48,35	8,87	1648/ 142= 11,6	184/ 1648= 10,93	83/ 562= 14,77	562/ 66= 8,51	12,21	562 / 1255= 44,78	26,29 19,6
Anxiety	29,38 26,23	1489/ 3267= 45,58	10,84	1489/ 166= 8,97	210/ 1489= 14,1	147/ 931= 15,79	931 / 104= 8,95	12,42	931/ 1954= 47,64	28,61 23,59
Constipation	26,59 23,93	2159/ 4709= 45,85	12,39	2159/ 244= 8,85	272/ 2159= 12,6	139/ 679= 20,47	679 / 77= 8,82	12,62	679/ 1457= 46,6	33,97 27,18
Delirium	35,58 30,17	759/ 1793= 42,33	7,69	759/ 76= 9,98	158/ 759= 20,82	98/ 339= 28,91	339 / 40= 8,47	12,67	339 / 749= 45,26	38,99 28,84
Fatigue	27,92 25,19	1987/ 4005= 49,61	11,11	1987/ 207= 9,6	252/ 1987= 12,68	155/ 1198= 12,94	1198 / 128= 9,36	11,03	1198/ 2390= 50,12	26,19 21,92
Fever	35,79 31,04	838/ 1917= 43,71	9,73	838/ 92= 9,11	167/ 8380= 19,93	116/ 512= 22,65	512 / 65= 7,87	14,07	512/ 1148= 44,59	35,37 27,79
Hypercalcemia	21,82 20,01	3025/ 7093= 42,65	8,94	3025/ 287= 10,54	293/ 3025= 9,68	159/ 990= 16,06	990 / 127= 7,79	9,28	990/ 2179= 45,43	28,09 23,36
Loss, grief, and bereavement	23,69 21,1	2871/ 6413= 44,77	9,83	2871/ 257= 11,17	295/ 2871= 10,27	199/ 2039= 9,76	2039 / 225= 9,06	9,92	2039/ 4683= 43,54	21,67 18,24
Lymphedema	32,71 29,4	1717/ 3571= 48,08	9,46	1717/ 154= 11,15	249/ 1717= 14,5	172/ 1075= 16	1075 / 121= 8,88	10,23	1075/ 2352= 45,7	29,08 24,02
Nausea and vomiting	22,97 21,08	2333 / 5541=	6,95	2333/ 198=	228/ 2333=	128/ 707=	707/ 76=	12,31	707/ 1452=	30,99 25,14

		42,1		11,78	9,77	18,10	9,3		48,69	
	Type/ token ratio(x100) (con / sin pal. forma)	**Palabras vinculadas (x100) / n. palabras**	**Coeficiente de cohesion**	**Palabras vinculadas/ n. de oraciones**	**Cat. conceptuales por 100 pal. vinculadas**	**Cat. conceptuales por 100 pal. vinculadas**	**Palabras vinculadas/ n. oraciones**	**Coeficiente de cohesion**	**Palabras vinculadas (x100) / n. palabras**	**Type/ token ratio(x100) (con/ sin pal. forma)**
Nutrition	27,53 25,15	2272/ 4609= 49,29	12	2272/ 185= 12,28	285/ 2272= 12,54	181/ 1406= 12,87	1406 / 159= 8,84	14,04	1406/ 2973= 47,29	26,20 22,23
Oral complications of cancer	28,73 26,42	2180/ 4651= 46,87	8,74	2180/ 181= 12,04	292/ 2180= 13,39	191/ 1415= 13,5	1415 / 164= 8,63	9,74	1415 / 3032= 46,66	24,14 20,09
Post-traumatic stress disorder	24,37 21,42	1824 / 3992= 45,69	5,96	1824/ 114= 16	210/ 1824= 11,51	149/ 911= 16,35	911 / 95= 9,58	9,44	911/ 1884= 48,35	28,77 23,3
Pruritus	30,82 28,16	1758/ 3761= 46,74	11,24	1758/ 193= 9,11	260/ 1758= 14,79	125/ 476= 26,26	476 / 72= 6,61	12,75	476/ 1030= 46,21	39,90 32,72
Sleep disorders	30,68 28,12	1554/ 3119= 49,82	12,43	1554/ 122= 12,74	194/ 1554= 12,48	98/ 323 = 30,34	323 / 23= 14,04	17,01	323 / 579= 55,78	45,08 35,58
Transitional care planning	29,15 25,47	1412/ 2792= 50,57	10,99	1412 / 141= 10,01	207/ 1412= 14,66	176/ 1530 = 11,5	1530 / 200= 7,65	10,43	1530/ 3683= 41,54	21,04 17,43
Desviación típica	5,51 4,55	2,63	1,83	1,88	4,77	5,89	1,6	2,03	3,15	6,66 5,13
Desviación típica ajustada	4,24 3,82	2,49			3,18	4,93			2,18	5,2 4,17

Al calcular la desviación típica de estos datos, se percibe que hay textos en los que los cálculos *Palabras vinculadas (x100)/ Nº total de palabras* o la *type/token ratio* difieren bastante con respecto al resto. Se han indicado estas anomalías mediante subrayado y se han eliminado en la media aritmética para no distorsionar los datos.

Entre las inferencias sobre nuestro corpus extraídas de las estadísticas del texto podemos citar que el número de palabras vinculadas por cada 100 palabras es ligeramente superior en textos para especialistas. En concreto, 46,55 frente al 45,86 correspondiente a textos para pacientes. Este porcentaje es alto por el carácter sintético de los resúmenes objeto de estudio (*Summaries for health professionals and patients*). La *type/token ratio* es elevada porque el corpus analizado no es muy extenso. Aún más elevada es la *standardised type/token ratio* por cada 300 palabras de texto.

En cada oración de QD, que tiene una extensión media de 22,15 palabras, hay 10,71 palabras vinculadas mientras que en QP, de las 17,71 palabras por oración, 8,69 están vinculadas. Aunque la proporción *palabras por oración/palabras vinculadas* es muy similar, si consideramos la cohesión como fenómeno interoracional, los textos para especialistas son más cohesivos porque a mayor número de palabras cohesionadas por oración, hay un mayor número de conexiones entre oraciones.

El número de categorías conceptuales activadas por cada 100 palabras vinculadas es menor en textos para especialistas, en parte porque las USE no son reconocidas por un tesauro de la lengua general. En los textos no especializados, con menos palabras se actualizan las mismas áreas conceptuales, aunque obviamente con menor profundidad. Por eso también, el coeficiente de cohesión es mayor.

11.3.4. Recapitulación

En definitiva, a partir de cálculos estadísticos básicos, se pueden extraer algunas inferencias sobre el tipo de relación cohesiva que se establece en diferentes tipos textuales:

a) Los datos aportados por el programa *Hesperus* a grandes rasgos no distan demasiado de los obtenidos por el analista, por lo que pueden aplicarse a un corpus más extenso.
b) La mitad de las unidades léxicas de los resúmenes analizados contribuyen a la cohesión del texto, aunque en los textos para pacientes son suficientes menos palabras cohesionadas para activar más áreas conceptuales.
c) La mayor extensión de las oraciones en textos para profesionales de la salud hace que haya más palabras vinculadas en cada oración con respecto a los textos para pacientes.
d) En definitiva, la estructuración conceptual que emana de un determinado campo del saber está relacionada con la cohesión de las cadenas léxicas generadas a partir de la activación textual de los conceptos. Por eso, las cadenas léxicas y la cohesión en textos sobre el mismo tema tiene que ser igual o muy parecida.