LINGÜÍSTICA DE CORPUS, ANÁLISIS DE REDES Y MATRICES DE CO-OCURRENCIAS

Keith Stuart, Ana Botella[1]
Universidad Politécnica de Valencia

ABSTRACT

En este artículo describimos la investigación que se ha desarrollado en el diseño de una metodología para la representación reticular del conocimiento que se genera en el seno de una institución a partir de un corpus representativo de la producción científica de los integrantes de dicha comunidad discursiva, la Universidad Politécnica de Valencia.. Para ello, presentamos las acciones que se realizaron en las fases iniciales del estudio encaminadas a establecer el marco teórico y práctico en el que se inscribe nuestro análisis. En la sección de metodología se describen las herramientas informáticas utilizadas, así como los procesos que nos permitieron disponer de aquellos elementos presentes en el corpus, que nos llevarían al desarrollo de matrices de co-ocurrencias con las que se generaron redes semánticas del conocimiento disciplinar. Finalmente, a partir de los resultados obtenidos, constatamos la viabilidad de extraer y representar el capital intelectual basándonos en los principios de la lingüística de corpus en combinación con las formulaciones de la teoría de redes.

Palabras clave: *lingüística de corpus, artículos académicos, matrices de co-ocurrencias, redes semánticas, descubrimiento del conocimiento*.

I. INTRODUCCIÓN

Este artículo propone un modelo para la aplicación del análisis de redes a la lingüística de corpus como método para la representación del conocimiento que se genera en nuestra comunidad discursiva académica. La idea de partida es sencilla: las palabras que conforman un corpus son los nodos de una red lingüística que los relaciona. El artículo analiza el discurso de la ciencia y de la tecnología mediante el estudio de palabras clave y su co-selección en artículos de investigación en un corpus de 1.376 artículos, todos ellos procedentes de revistas especializadas (un total de 6.104.323 palabras). Todos los artículos que componen el corpus han sido escritos por nuestros profesores e investigadores y representan el trabajo de una única comunidad discursiva, la Universidad Politécnica de Valencia. Dichos artículos se han publicado en revistas indexadas en el *Science Ctation Index* (*SCI*®).

La hipótesis primera de la que partimos en nuestra investigación es demostrar cómo el lenguaje, y en este caso el texto escrito, es el vehículo de intercambio y transmisión del conocimiento entre los miembros de una comunidad discursiva. Se trata, pues, de extraer el conocimiento que ha quedado plasmado en artículos científicos, analizarlo y organizarlo para poderlo representar. Para ello, partimos del corpus seleccionado y su posterior análisis, lo cual nos permitirá determinar a través de la terminología utilizada, el estudio microscópico y

[1] **Address for correspondence:** Departamento de Lingüística Aplicada. Universidad Politécnica de Valencia. 46022 Valencia. Tel.:96-652-8495. *kstuart@idm.upv.es*

macroscópico de ciertos rasgos léxico–gramaticales cuál es ese conocimiento que se genera en el contexto universitario.

Según la tradición firthiana, las colocaciones manifiestan ciertas afinidades léxicas y semánticas que van más allá de las restricciones gramaticales. Sinclair (1991:170) se refiere a la colocación como "the occurrence of two or more words within a short space of each other in a text"; esta definición podría lógicamente hacer referencia a la co-selección entre ítems léxicos o gramaticales. Desde la perspectiva de las redes, podemos plantear el concepto de la siguiente manera: si dos unidades *a, b* están relacionadas en términos de estadísticas colocacionales (o simplemente como bigramas frecuentes) como lo están las unidades *b*, *c*, hay implícita una relación indirecta, incluso aunque no esté confirmada directamente por una colocación de *a* y *c*. Hemos sido cautos en el estudio en nuestras asunciones, utilizando sólo los colocativos / bigramas de aquellas palabras que se habían obtenido como palabras clave y considerando como nodos cohesivos solamente aquellas palabras clave que estaban relacionadas al menos tres veces (Hoey, 1991).

El artículo explica cómo generamos matrices de co-ocurrencias de palabras clave y cómo visualizamos la co-aparición de dichas palabras clave en 23 áreas diferentes de conocimiento especializado y en el corpus en su totalidad. Para esta tarea, tuvimos que utilizar diferentes programas informáticos. Se usó *Wordsmith* para extraer las palabras clave del corpus. Se obtuvo un listado inicial de palabras clave al comparar nuestro corpus de inglés científico, que hemos llamado Corpus UPV, con un corpus de inglés general (el *British National Corpus*). A su vez, se extrajeron listados de las palabras clave de cada uno de los 23 dominios especializados al comparar ese listado inicial de palabras clave del corpus UPV con cada una de dichas áreas (*key-key words*). Las matrices de palabras clave se confeccionaron mediante códigos de desarrollo propio en *Perl* y se volcaron a hojas de cálculo. En este punto, se transfirió cada una de ellas al programa *Ucinet* y, finalmente, se visualizaron las redes con la utilidad *Netdraw*.

El objetivo prioritario del artículo es mostrar cómo esas redes intratextuales e intertextuales generadas a partir de las palabras clave nos ofrecen gránulos de fragmentos del conocimiento que se encuentra disperso en y a través de los textos, y que contienen una elevada carga semántica.

Los avances en teoría de redes no sólo proporcionan un marco adecuado de integración, sino que abrirán nuevas perspectivas en el estudio del lenguaje y la organización del conocimiento. En este sentido, la lingüística de corpus en combinación con el análisis de redes puede convertirse en una técnica aplicable al descubrimiento del conocimiento y, en nuestro caso particular, el conocimiento disciplinar.

II. MÉTODO

En el estudio hemos podido descubrir cómo las palabras dependientes del dominio, los términos, presentan estadísticas de baja frecuencia en la interacción normal de la población en general, pero no ocurre lo mismo en los contextos especializados, donde éstos se han

engendrado. Los términos ayudan a definir las comunidades que los utilizan en la misma medida en que estas comunidades definen sus términos. La información recopilada en las diferentes etapas de la investigación se ha dispuesto siguiendo las nociones de frecuencia de palabras, palabras clave y relaciones léxicas y gramaticales, es decir, aquellos fenómenos conocidos como *collocation, semantic prosody* y *colligation*. Asimismo, basándonos en la relevancia estadística, se ha valorado el grado de interacción, las asociaciones, que se producen entre determinados ítems relevantes a nuestra investigación.

Además del estudio intratextual, se han abordado ciertos aspectos intertextuales que nos han permitido detectar las variaciones que se producen dentro del mismo género. Para todo ello, hemos trabajado en el desarrollo de determinadas aplicaciones informáticas diseñadas a medida conforme a nuestras necesidades y hemos podido comparar tanto las ventajas y las debilidades de dichas herramientas como los resultados obtenidos tras su aplicación con otras existentes en el mercado para fines similares, como por ejemplo *Wordsmith Tools*. Dichas cuestiones se han contemplado tanto en el ámbito de nuestra institución, de forma global, como en cada uno de los dominios especializados de forma particular.

Una vez concluido el análisis intratextual, en la etapa posterior se llevó a cabo un análisis que nos permitiese cuantificar y representar aspectos concretos sobre la variación y recursividad a nivel intertextual. Partimos de la premisa de que por encima del nivel de textos individuales existe una estructura de formación a la que se puede acceder mediante la lingüística de corpus.

Autores como Kristeva (1966), Barthes (1970) o Bajtín (1986) entienden la intertextualidad en el sentido de que un texto siempre está vinculado a textos o experiencias previas y muestran prospección a textos o enunciados futuros. De Beaugrande y Dressler (1981) afirman que cualquier texto debe cumplir con el requisito de la intertextualidad para que pueda considerarse como texto, que, además, determina la manera en que el uso de un cierto texto depende del conocimiento de otros textos. Para estos autores, el término intertextualidad se refiere a la relación de dependencia que se establece entre, por un lado, los procesos de producción y recepción de un texto determinado y, por otro, el conocimiento que tengan los participantes en la interacción comunicativa de otros textos anteriores relacionados con él.

Según Stubbs (1996), los textos se orientan conforme a rutinas y convenciones; están modelados por textos previos a los que hacen referencias intertextuales, probablemente incluidas en el mismo corpus. En este sentido, Stubbs (2001: 120) señala, "Analysis cannot be restricted to isolated texts. It requires an analysis of intertextual relations, and therefore comparison of individual instances in a given text, typical occurrences in other texts from the same text-type, and norms of usage in the language in general".

En esta misma línea, encontramos la postura de Fairclough (2002), quien defiende una perspectiva intertextual para el análisis, por ejemplo, de frases pre-construidas y colocaciones fijas.

Una vez delimitado el marco en el que se inscribe esta fase del estudio, definimos como objetivos específicos:

- Representar la frecuencia de cada palabra clave en cada unos de los diferentes documentos que componen las áreas de conocimiento del Corpus UPV
- Representar la distribución de cada una de esas palabras claves en las diferentes secciones características del artículo de investigación (*IMRD*)
- Relacionar y representar las interacciones entre los términos conforme a su índice de frecuencias
- Comparar y representar el grado de recursividad que se produce en cuanto a patrones idénticos de diferente longitud (*clusters*) en cada uno de los textos analizados

El trabajo se realizó en cuatro etapas sucesivas que se concretan en la siguiente tabla:

Tabla 1. Generación de matrices

Generación de matrices: Análisis intratextual e intertextual a partir de los procedimientos anteriores
Matriz 1: Distribución de palabras clave por documento
Matriz 2: Distribución de palabras clave por sección del artículo
Matriz 3: Combinaciones entre palabras clave
Matriz 4: Distribución de *clusters* (de 3 a 8 palabras) por documento

El esquema básico que se siguió para cada una de las matrices fue el que a continuación se ofrece:

Tabla 2. Esquema generación de matrices

Matriz 1

	Doc 1	Doc 2	Doc 140
Palabra 1	Frequencia		
Palabra 2			
Palabra 3			
Palabra *n*			

Matriz 2

	Abstract	*Introduction*	*Methods*	*Results*	*Discussion*	*Conclusion*
Pal. 1	Freq.					
Pal. 2						
Pal. 3						
Pal.*n*						

Matriz 3

	Result	*System*	Palabra *3*			*Palabra 100*
Palabra 1	Frequencia					
Palabra 2						
Palabra 3						
Palabra *n*						

Matriz 4

	Doc 1	Doc 2	Doc 140
Clusters 3 palabras	Frequencia		
Clusters 4 palabras			
Clusters n = 8			

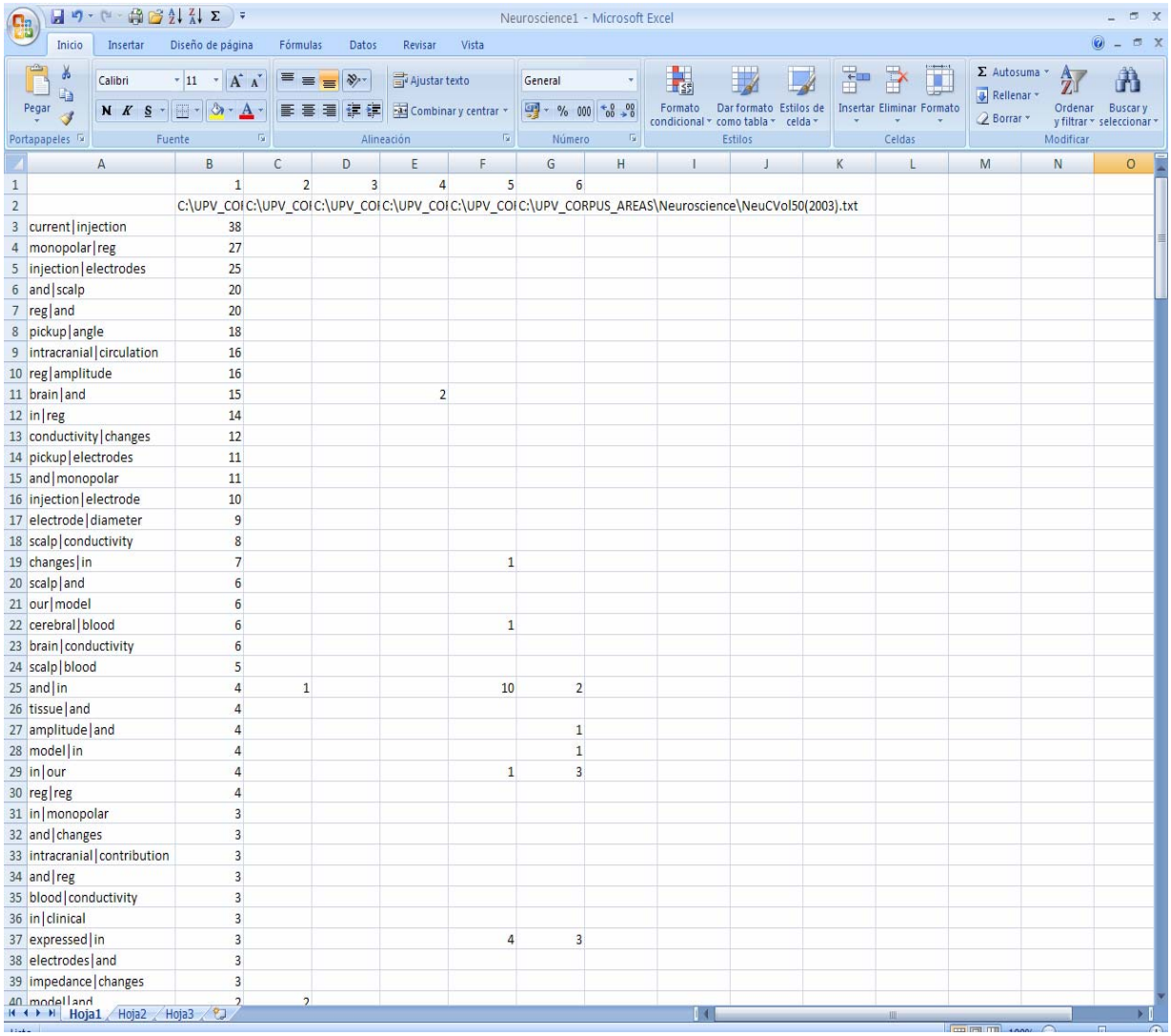
Las matrices se generaron a partir de los listados de *key-words* de cada área de conocimiento y también del corpus en su totalidad de forma global. Se utilizó para ello una aplicación de desarrollo propio en lenguaje *Perl* y se transfirió cada una de ellas a una hoja de cálculo.

La siguiente fase consistió en valorar y determinar qué tipo de herramienta informática podría ser la adecuada para llevar a cabo la representación de la información en forma reticular a partir de determinados aquellos aspectos intratextuales e intertextuales que se habían dispuesto en forma de matriz. La herramienta *Ucinet 6* demostró reunir las condiciones para tales fines. Por ello, utilizando la utilidad *Netdraw* de la herramienta, se procedió a realizar diferentes representaciones que nos permitieron llegar a establecer conclusiones sobre la representación gráfica del conocimiento a partir de las matrices de co-ocurrencias de palabras clave.

Ucinet es una herramienta para la representación de redes sociales. El análisis de redes sociales constituye un método de valoración de redes informales mediante la representación de las relaciones entre personas, equipos, departamentos o incluso organizaciones enteras. Estudia la forma en que individuos u organizaciones se conectan y define la posición que éstos ocupan en la red, sus grupos y estructura global, los flujos de conocimiento e información y las relaciones de influencia recíproca. Este tipo de análisis se ha venido aplicando desde hace algún tiempo para investigar la colaboración entre autores o instituciones en publicaciones científicas. Ejemplos de este tipo de iniciativa los encontramos en De Solla (1965), Kretschmer (1994), Melin y Persson (1996), Newman (2001), Molina y Muñoz (2002), Sanz (2003), de Granda et al. (2005), González Alcaide et al. (2006).

III. RESULTADOS

En la primera de dichas matrices se representa la organización de los pares de palabras clave en cada unos de los documentos que se encuentran en cada área del Corpus UPV. De esta forma, podemos identificar aquellos pares que son específicos a un artículo concreto frente aquellos que se repiten en más de un documento. La matriz que hemos tomado como ejemplo corresponde al área de *Neuroscience*. Al tratarse de un dominio con un número reducido de artículos, podemos ofrecer una impresión de pantalla en la que se puede apreciar la disposición de los elementos en la hoja.



	A	B	C	D	E	F	G
1		1	2	3	4	5	6
2		C:\UPV_COI	C:\UPV_COI	C:\UPV_COI	C:\UPV_COI	C:\UPV_COI	C:\UPV_CORPUS_AREAS\Neuroscience\NeuCVol50(2003).txt
3	current|injection	38					
4	monopolar|reg	27					
5	injection|electrodes	25					
6	and|scalp	20					
7	reg|and	20					
8	pickup|angle	18					
9	intracranial|circulation	16					
10	reg|amplitude	16					
11	brain|and	15			2		
12	in|reg	14					
13	conductivity|changes	12					
14	pickup|electrodes	11					
15	and|monopolar	11					
16	injection|electrode	10					
17	electrode|diameter	9					
18	scalp|conductivity	8					
19	changes|in	7				1	
20	scalp|and	6					
21	our|model	6					
22	cerebral|blood	6				1	
23	brain|conductivity	6					
24	scalp|blood	5					
25	and|in	4	1			10	2
26	tissue|and	4					
27	amplitude|and	4					1
28	model|in	4					1
29	in|our	4				1	3
30	reg|reg	4					
31	in|monopolar	3					
32	and|changes	3					
33	intracranial|contribution	3					
34	and|reg	3					
35	blood|conductivity	3					
36	in|clinical	3					
37	expressed|in	3				4	3
38	electrodes|and	3					
39	impedance|changes	3					
40	model|and	2	2				

Figura 1. Pantalla Matriz 1: *Bi-grams* por documento

La red que a continuación presentamos refleja cómo la mayoría de los *bi-grams* suelen agruparse en torno a cada uno de los documentos, si bien algunos de ellos, con menor carga semántica, se encuentran en posiciones intermedias por ser comunes a más de un documento.

Podemos tras este primer análisis afirmar que los pares con una densidad semántica elevada suelen concentrarse en un texto, lo cual denota la especificidad de cada uno de los artículos analizados.

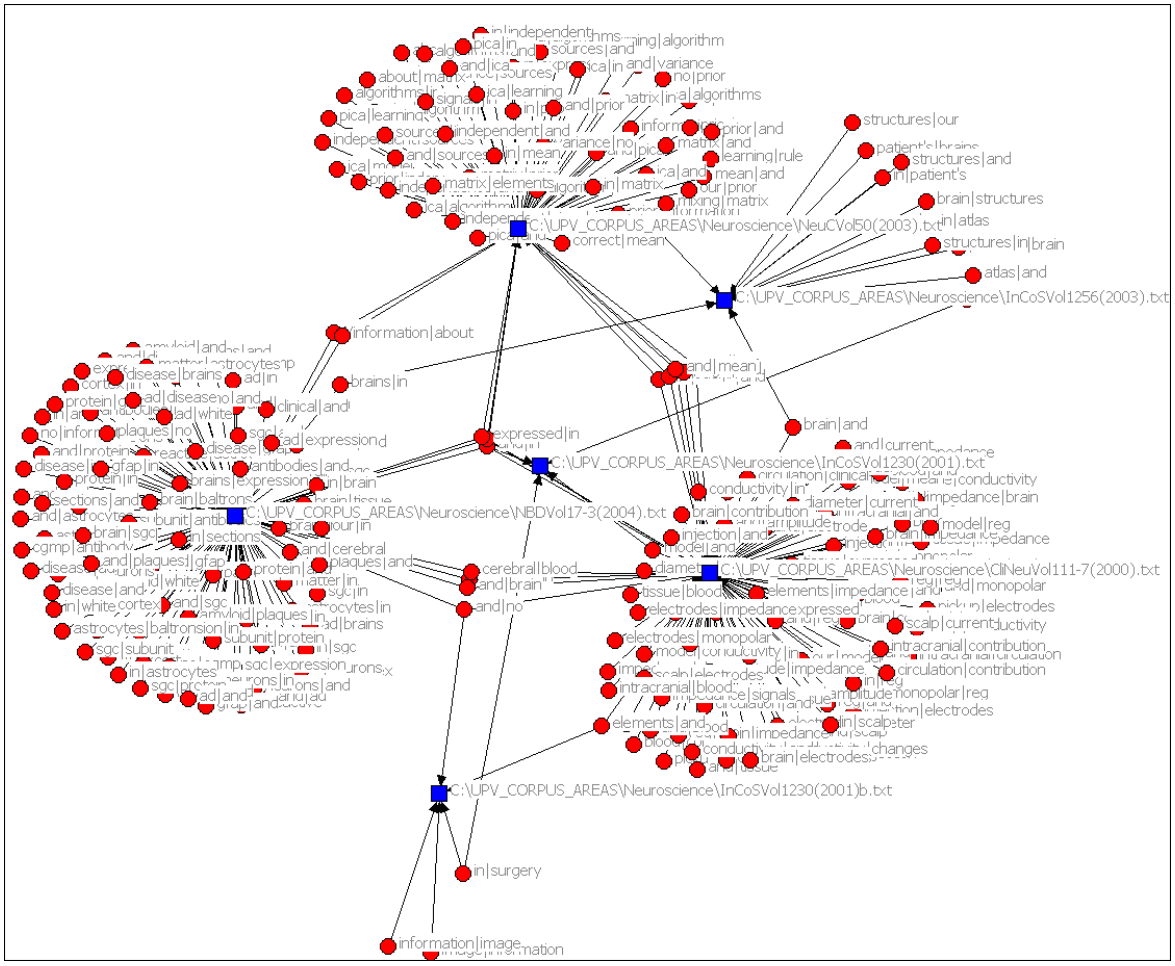


Figura 2. Ejemplo de red de *Bi-grams* por documento

La siguiente matriz generada representa la distribución de las palabras clave en cada una de las secciones características del artículo académico *(Abstract, Introduction, Methods, Results, Discussion, Conclusion).*

La información que nos proporciona resulta ser un indicativo claro del *aboutness* de los textos que componen los diferentes subdominios. Si al analizar los listados de palabras clave de las diferentes áreas en etapas previas anteriores de nuestro estudio, obtuvimos información sobre el conocimiento implícito de forma global, en la presente fase, disponemos de las herramientas necesarias para interpretar cuantitativamente cómo se estructura esa información léxica en las secciones estándar de los documentos. Este tema ha sido objeto de estudio por parte de diferentes autores, de áreas también diversas, que se basan en la minería de textos para descubrir el conocimiento que se encuentra en un elevado numero de textos y que sería imposible averiguar de forma manual mediante una lectura exhaustiva. Resulta interesante

destacar que la mayoría de estudios han basado su análisis solamente procesando la sección de *abstracts* de los artículos. Un análisis similar, a partir de las palabras clave en las diferentes secciones de artículos académicos, es el realizado por Shah et al. 2003. El motivo para centrarse en dicha sección responde, por una parte, a la disponibilidad de los resúmenes on-line y, por otra, a la elevada cantidad de información que en éstos se concentra. No obstante, la sección de resultados es la que cubre más información del artículo, mientras que la de resúmenes contienen la mayor densidad de información (Schuemie et al., 2004).

Si observamos la siguiente tabla, tomada del área de *Chemistry*, que muestra la distribución de palabras clave en cada una de las secciones, encontramos que en los *Abstracts* se repiten de forma significativa (teniendo en cuenta que esta parte es más reducida en extensión) términos como *compound, polymeric, immunoassay* y *pesticides*. En *Methods* y *Results* son relevantes términos como: *curves, fig., observed, concentration/s, range, calibration*, que se utilizan para expresar los hallazgos tras un proceso o modelo de investigación. La información que obtenemos por el procedimiento del análisis por secciones nos lleva a concluir que existe una determinada preferencia o concentración en el uso de ciertos términos en las diferentes partes del artículo académico. Podemos, incluso, afirmar que éstos se pueden agrupar bajo categorías dado que suelen mostrar rasgos comunes.

Tabla 3. Distribución de 50 palabras clave por secciones del documento *(Chemistry)*

Palabra	**Abstract**	**Introduction**	**Methods**	**Results**	**Conclusion**
1. temperature	284	430	386	310	315
2. peak	47	131	226	218	122
3. potential	86	125	156	172	117
4. sample	120	264	230	221	205
5. curves	21	81	**148**	**149**	82
6. water	233	228	234	251	206
7. ph	70	139	170	164	108
8. peaks	30	33	109	105	65
9. fluorescence	46	51	55	56	50
10. elisa	31	30	39	46	51
11. presence	113	94	124	130	132
12. compounds	**100**	55	49	80	83
13. compound	45	63	47	62	80
14. fig	70	411	**742**	**752**	440
15. determination	77	50	41	26	70
16. antibody	19	28	53	21	32
17. curve	16	39	85	94	66
18. acid	139	126	121	126	121
19. experiments	82	**164**	87	83	87

Palabra	Abstract	Introduction	Methods	Results	Conclusion
20. chemical	116	76	56	64	68
21. organic	88	40	47	44	50
22. assay	24	28	47	51	37
23. chimica	36	19	26	23	25
24. experimental	119	145	145	141	122
25. found	94	82	133	116	181
26. solution	148	301	273	206	226
27. observed	90	123	**191**	**214**	178
28. interaction	75	27	45	59	77
29. polymeric	**35**	15	12	14	37
30. immunosensors	22	21	10	9	24
31. immunosensor	25	10	14	15	28
32. solvents	33	29	34	23	29
33. concentration	78	108	**118**	**150**	128
34. range	79	96	**100**	**118**	79
35. samples	104	184	168	132	210
36. liquid	66	71	39	42	43
37. solutions	82	153	85	74	52
38. mobility	39	11	19	16	14
39. adsorbed	22	26	31	22	25
40. reported	75	66	68	75	78
41. prepared	60	111	37	19	32
42. pesticide	25	14	12	10	11
43. immunoassays	**29**	12	10	6	10
44. measured	63	114	**120**	89	52
45. immunoassay	25	4	9	17	18
46. buffer	17	58	73	46	23
47. binding	**62**	27	18	46	26
48. pesticides	**42**	9	7	5	11
49. concentrations	22	47	**74**	**77**	32
50. calibration	14	27	**31**	**26**	26

La distribución de los términos recogida en forma de matriz puede visualizarse utilizando la herramienta *Netdraw*. Al pinchar en cada unos de los términos, veremos el número de aristas que conectan con las diferentes categorías, en este caso, las secciones de los textos. De esta manera apreciamos cómo un término está contenido en una o más secciones.

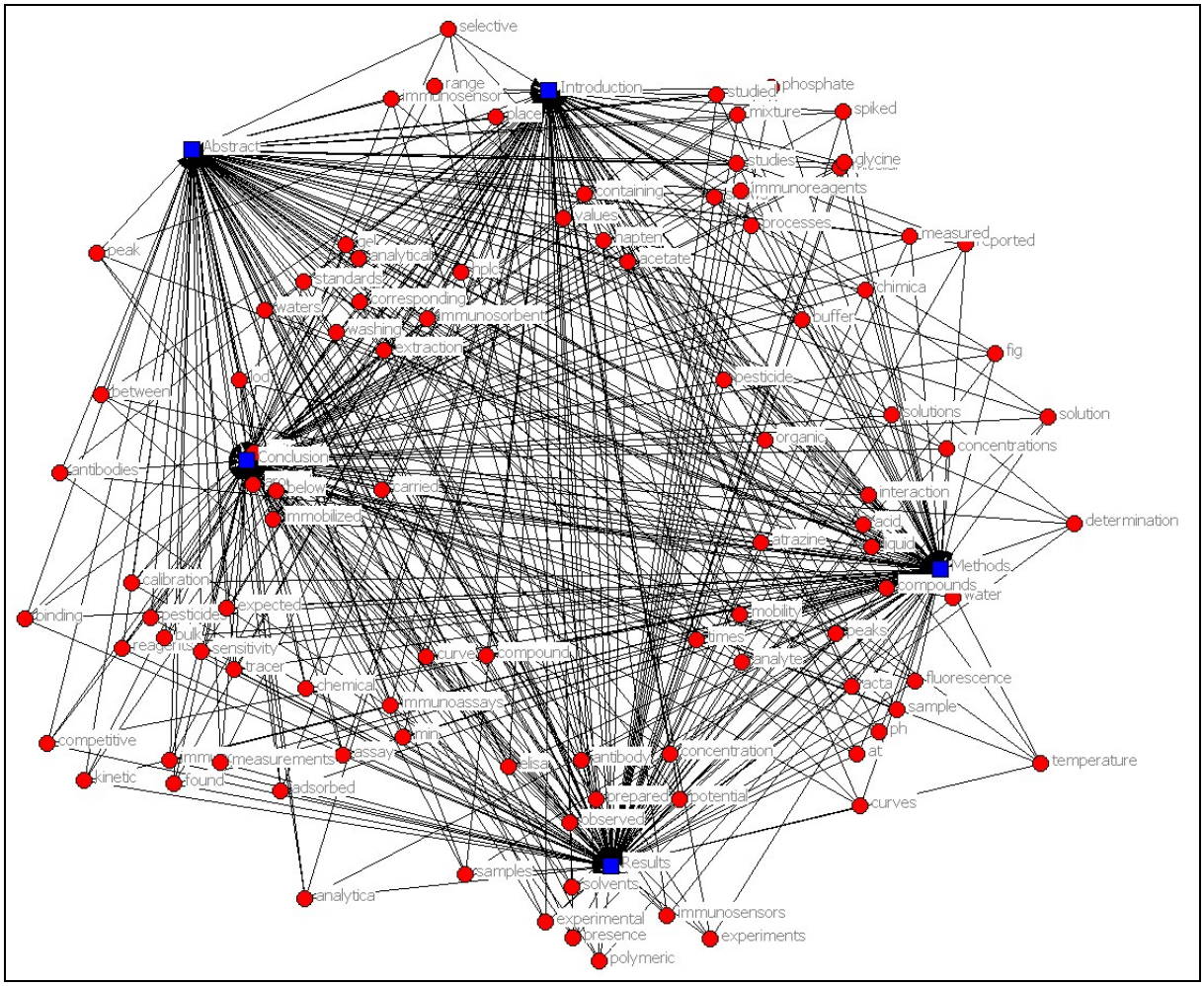
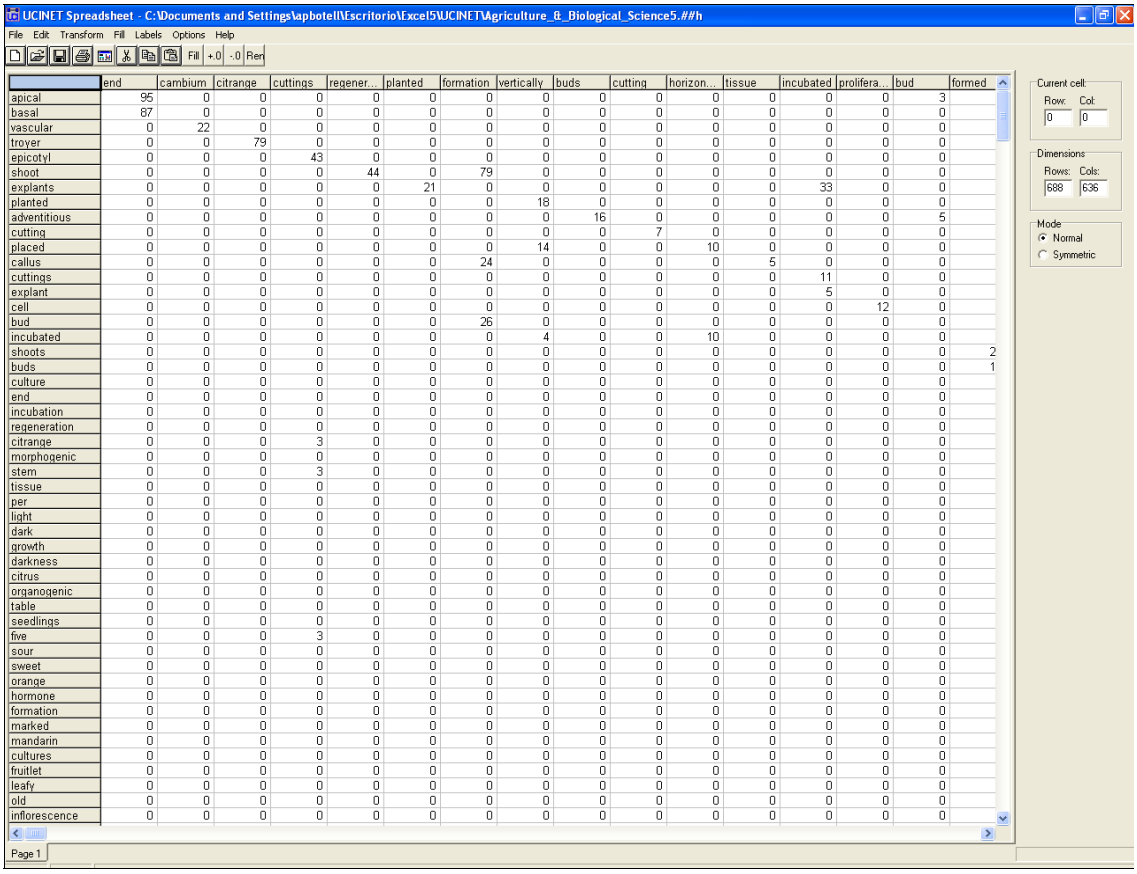


Figura 3. Ejemplo de red de palabras clave por secciones del documento

En la Matriz 3 se han recogido las combinaciones entre palabras clave *(bi-grams)*. El objetivo de este tipo de análisis es determinar cómo un mismo elemento se relaciona con mayor o menor frecuencia con otros elementos relevantes del dominio. Por este motivo, la matriz está formada en ambos ejes, vertical y horizontal, por los mismos elementos, las palabras claves de cada área, y en las celdas figura el número de combinaciones que se producen entre ellas. La disposición de la información de esta manera permite detectar si la combinación entre un par es unidireccional o bidireccional, así como el número de repeticiones. Por ejemplo, comprobamos que la combinación *apical end* es muy frecuente (95 repeticiones), como también lo es *basal end* (87 apariciones), siendo mucho más reducido el uso de la combinación *apical bud* (3 casos). Ahora bien, en este punto nos preguntamos:

1. ¿son las combinaciones *apical end* y *basal end* unidireccionales o bidireccionales?
2. ¿con qué otro/s término/s se relaciona el término *apical*?
3. ¿con qué otro/s términos se combinan *end* y *basal*?



	end	cambium	citrange	cuttings	regener...	planted	formation	vertically	buds	cutting	horizon...	tissue	incubated	prolifera...	bud	formed
apical	95	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
basal	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
vascular	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	
troyer	0	0	79	0	0	0	0	0	0	0	0	0	0	0	0	
epicotyl	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	
shoot	0	0	0	0	44	0	79	0	0	0	0	0	0	0	0	
explants	0	0	0	0	0	21	0	0	0	0	0	0	33	0	0	
planted	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	
adventitious	0	0	0	0	0	0	0	0	16	0	0	0	0	0	5	
cutting	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	
placed	0	0	0	0	0	0	0	14	0	0	10	0	0	0	0	
callus	0	0	0	0	0	0	24	0	0	0	0	5	0	0	0	
cuttings	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	
explant	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	
cell	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	
bud	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	
incubated	0	0	0	0	0	0	0	4	0	0	10	0	0	0	0	
shoots	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
buds	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
culture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
end	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
incubation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
regeneration	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
citrange	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	
morphogenic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
stem	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	
tissue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
per	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
light	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
dark	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
growth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
darkness	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
citrus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
organogenic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
table	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
seedlings	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
five	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	
sour	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
sweet	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
hormone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
formation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
marked	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
mandarin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
cultures	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
fruitlet	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
leafy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
old	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
inflorescence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figura 4. Ejemplo de matriz de combinaciones entre palabras clave (*Agriculture & Biological Sciences*)

Al comprobar en la tabla, observamos que *apical end* se utiliza solamente en este sentido, mientras que *basal end* es bidireccional, aunque el par *end basal* es menos frecuente (3 repeticiones). En respuesta a la segunda pregunta, vemos que *apical* se relaciona también con *bud* y con *shoot*. A su vez, encontramos la combinación *adventitious bud* y *bud formation*. Podríamos de este modo ir extendiendo las relaciones.

De forma similar, al buscar en la matriz las relaciones de *end* con otros términos, encontramos los ejemplos como *end table*, *stylar end*. Encontramos que *stylar* no se presenta en combinación con otros términos. *Basal*, a su vez se combina con *medium* (3 repeticiones), con *diet* (20 casos).

Al tomar como ejemplo un área con un elevado número de muestras, la matriz que se obtiene es también de grandes dimensiones, por lo que al intentar representar de forma gráfica el contenido de la misma, descubrimos que se trata de una red compleja, lo cual denota la complejidad del lenguaje, en la que podemos visualizar todos los vínculos existentes.

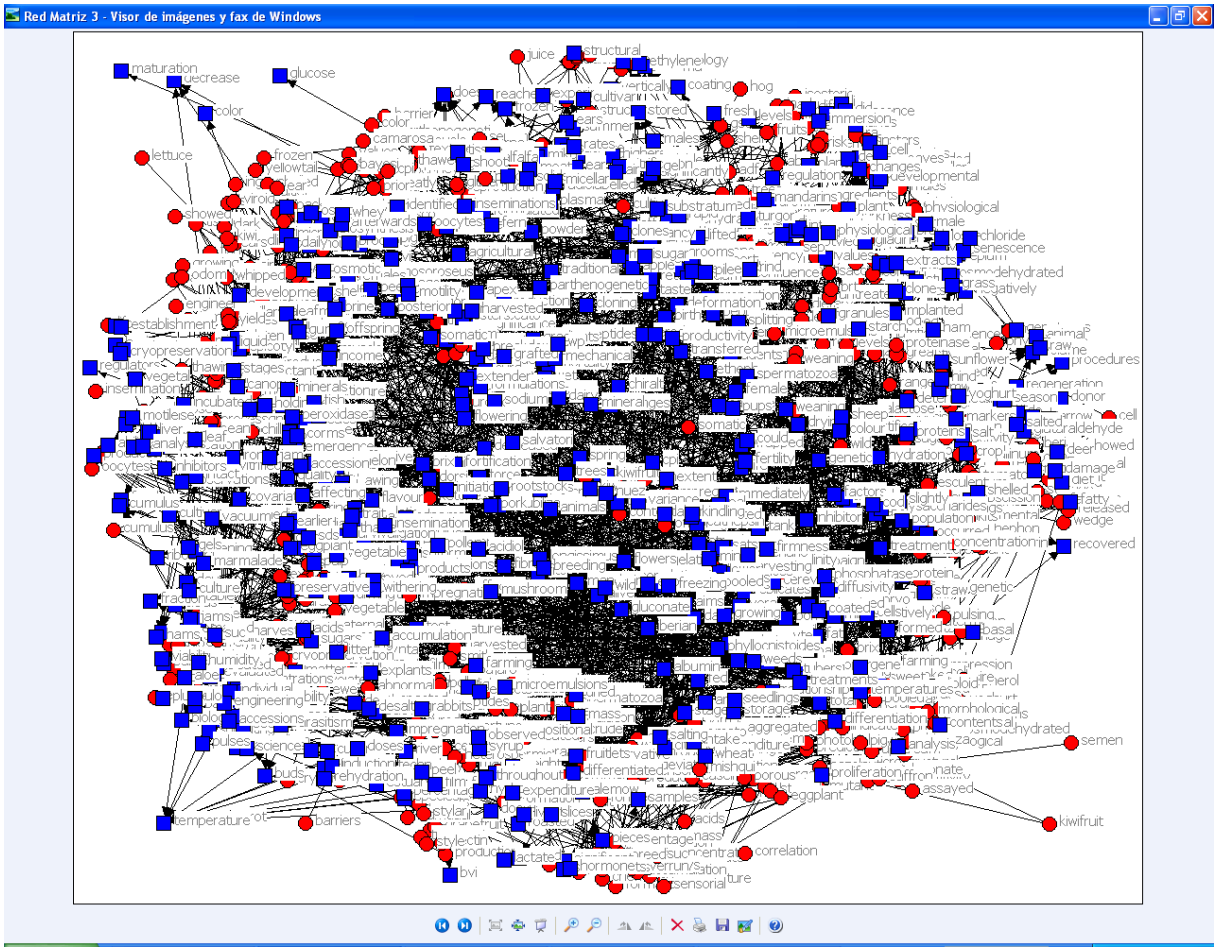


Figura 5. Ejemplo de red de combinaciones entre palabras clave (*Agriculture & Biological Sciences*)

En esta maraña de interacciones existe la opción de aplicar un filtro con un número mínimo de apariciones, de forma que las líneas por debajo del número establecido quedarán transparentes, si bien, también es posible mediante la opción *ego* en la barra de herramientas, situarnos sobre uno de los elementos, y visualizar únicamente las relaciones que dicho participante de la red presenta.

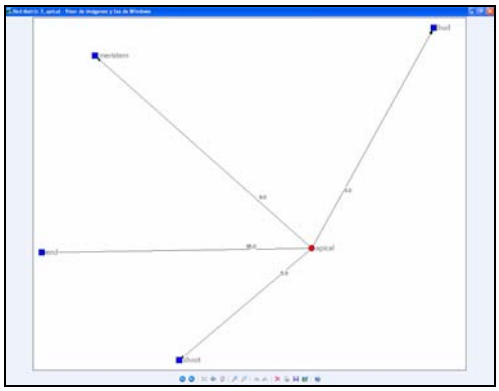


Figura 6. Ejemplo de red de combinaciones entre palabras clave a partir de 1 término (*Agriculture & Biological Sciences*)

Igualmente, la aplicación permite realizar búsquedas múltiples, seleccionando los elementos deseados, por ejemplo las *-n* palabras clave más frecuentes, y representar sus relaciones. A continuación presentamos la red generada a partir de los 10 primeros términos en la matriz.

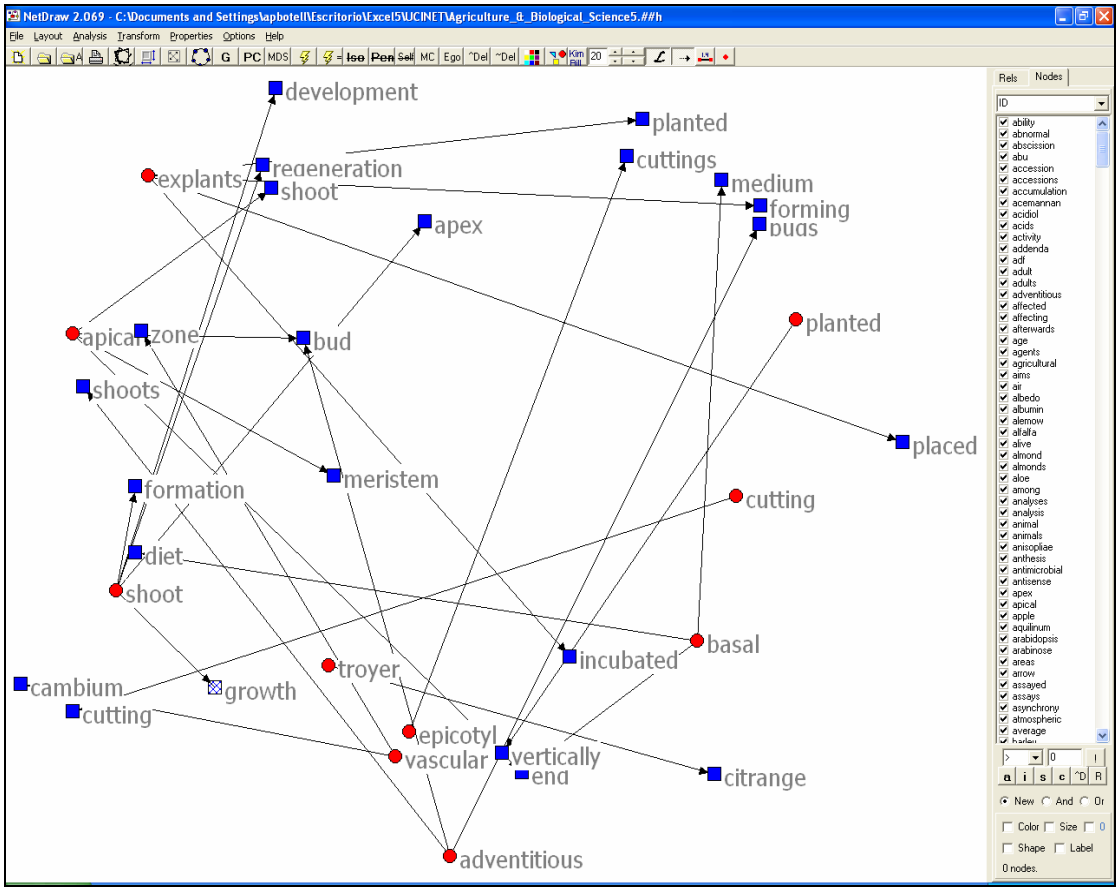


Figura 7. Ejemplo de red de combinaciones entre palabras clave a partir de 10 términos (*Agriculture & Biological Sciences*)

Ahora bien, el investigador deberá decidir qué es lo que le interesa según su punto de mira: por una parte puede visualizar todas las relaciones que un término presenta, con lo que obtendrá una dispersión conceptual, es decir, cubrirá de forma extensiva los diferentes conceptos en los documentos sometidos al análisis. Por otra parte, también podrá centrarse tan sólo en aquellas combinaciones que son más fuertes, esto es, más frecuentes, y por tanto albergan en sí una mayor densidad conceptual.

Nos encontramos ante *redes sociales* del lenguaje en las que los individuos o actores no son los miembros de un colectivo, sino términos y las aristas son las relaciones entre ellos. Metafóricamente hablando, de igual manera en que en las redes sociales, manejamos consideraciones en cuanto al tipo de relación entre los individuos: el número de veces que éstos se encuentren con determinados usuarios del sistema implicará una relación más o menos estrecha o significativa (densidad conceptual). Por el contrario, el número total de

participantes con que un mismo actor se encuentre, sin considerar si existen encuentros más o menos frecuentes, implicará una mayor complejidad en la red, aunque quizá con vínculos menos consistentes. Nos encontramos, pues ante un entramado léxico que nos ha permitido generar mapas del conocimiento explícito de una comunidad académica.

La siguiente matriz, Matriz 4, recoge las *clusters* o cúmulos (cadenas que van desde 3 hasta 8 palabras) de expresiones en cada uno de los artículos de los diferentes dominios del Corpus UPV, y también del Corpus de forma global, pauta que hemos seguido en las diferentes etapas de nuestro estudio.

Tabla 4. Ejemplo de distribución de *clusters* (3 palabras) por documento (*Agriculture & Biological Sciences*)

	ABVol84-6(1999).txt	ABVol85-1(2000).txt	ABVol86-1(2000).txt	ABVol87-6(2001).txt	AE&EVol 95-1(2003).txt
IN ORDER TO	0	0	0	0	0
THE EFFECT OF	2	8	9	1	1
THE NUMBER OF	13	15	16	12	0
DUE TO THE	1	3	0	1	1
THE END OF	2	4	0	10	0
END OF THE	20	6	11	1	0
THE PRESENCE OF	2	11	17	1	0
THE USE OF	0	0	1	0	1
A FUNCTION OF	0	0	0	0	0
WAS CARRIED OUT	0	0	0	0	0
AT THE END	2	4	0	6	0
THE INFLUENCE OF	3	5	4	3	0
AS A FUNCTION	0	0	0	0	0
CAN BE OBSERVED	0	0	0	0	0
ON THE OTHER	1	0	0	1	0
THE OTHER HAND	1	0	0	1	0
ACCORDING TO THE	1	2	1	0	4
CHANGES IN THE	0	0	0	2	0
EFFECT OF THE	0	3	3	0	8
THE PERCENTAGE OF	0	4	1	4	0
ARE SHOWN IN	0	0	0	0	0
IN TERMS OF	0	0	0	0	0
RELATED TO THE	1	0	0	3	0

Al analizar en fases anteriores los términos en el corpus UPV mediante la extracción de las cadenas recurrentes, pudimos comprobar que, dependiendo del horizonte que fijemos, obtendremos cadenas con rasgos diferentes. En secuencias de corta extensión, como la que presentamos en la tabla anterior, detectamos expresiones que son comunes a las diferentes áreas, por tratarse de expresiones frecuentes en los artículos académicos. En la mayor parte de los casos, son patrones que se repetirán en la mayoría de los textos. La siguiente red nos facilita la visualización de este aspecto:

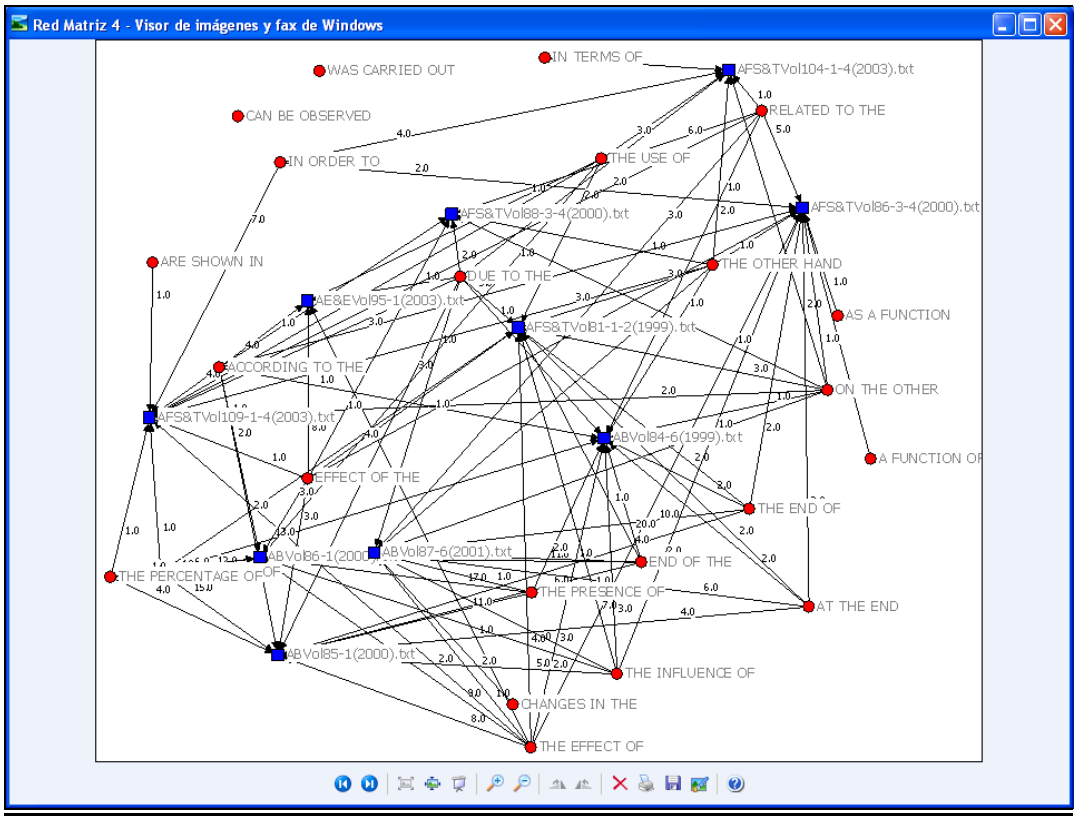


Figura 8. Ejemplo de red de *clusters* (3 palabras) por documento (*Agriculture & Biological Sciences*)

No obstante, a medida que las cadenas adquieren mayor longitud, detectamos que nos ofrecen mayor información conceptual.

Tabla 5. Ejemplo de distribución de *clusters* (6 palabras) por documento *(Agriculture & Biological Sciences*)

embryo recovery and in vitro development
embryos recovered in does with at
for growth rate from weaning to
for r and v lines respectively
for the explants incubated in the
from birth to the first week
from the marginal posterior density b
gold coated and viewed in the

growth rate from weaning to slaughter
had a significant effect on the

La red resultante de este tipo de análisis demuestra que estos *clusters*, al contener mayor información conceptual, son más característicos de determinados artículos.

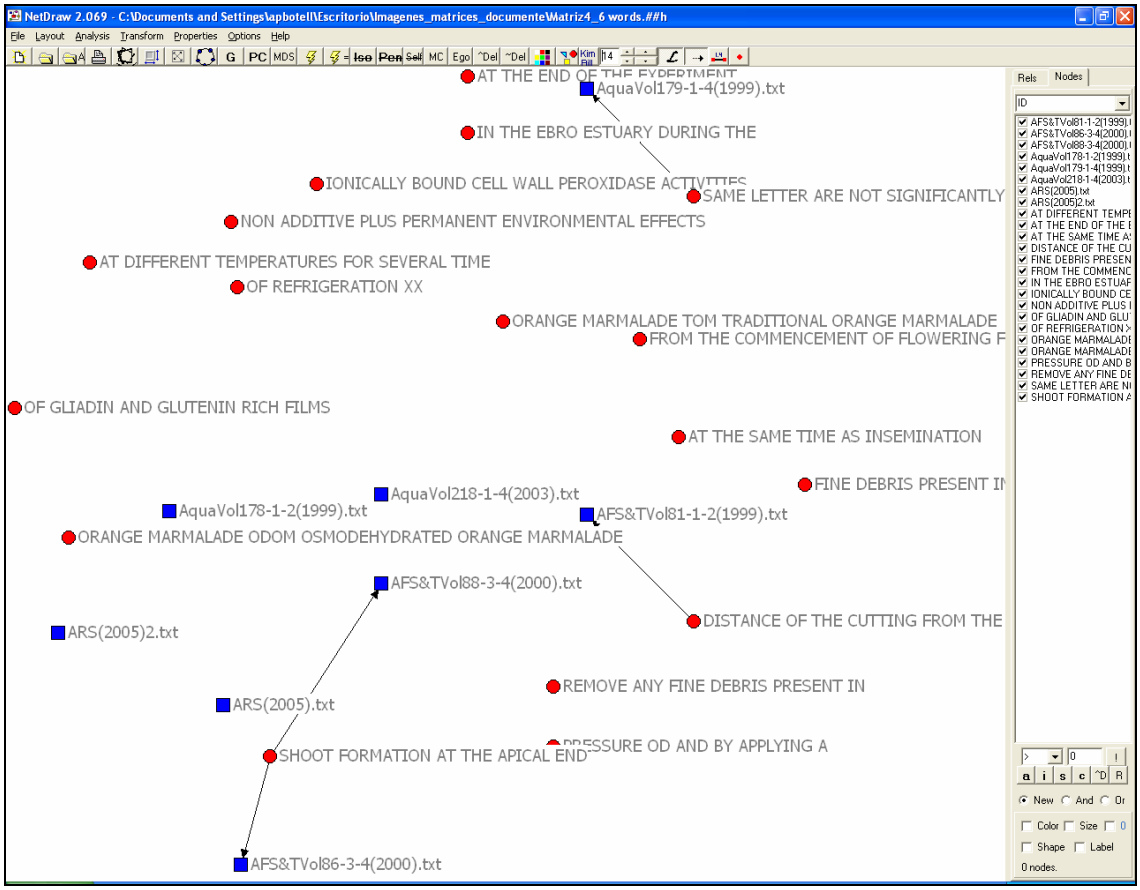


Figura 9. Ejemplo de red de distribución de *clusters* (6 palabras) por documento (*Agriculture & Biological Sciences*)

Concluimos nuestra exposición de los resultados obtenidos tras la aplicación metodológica que proponemos con una cita que confirma nuestras afirmaciones sobre el carácter complejo del lenguaje: "Language is clearly an example of a complex dynamical system. It exhibits highly intricate network structures at all levels (phonetic, lexical, syntactic, semantic) and this structure is to some extent shaped and reshaped by millions of language users over long periods of time, as they adapt and change them to their needs as part of ongoing local interactions" (Solé et al., 2005: 3).

IV. CONCLUSIONES

En este intento de esbozar un modelo para la generación de redes semánticas basándonos en la idea de las redes sociales (Barabási, 2002; Barabási & Jeong, 2002), y haciendo uso de

herramientas informáticas desarrolladas para tal fin (Borgatti, 2003), hemos llevado a cabo nuestro estudio tomando como referente los principios de la lingüística de corpus, metodología empírica que ha demostrado ser un procedimiento adecuado para llegar a disponer de la información necesaria sobre el lenguaje y sobre el conocimiento.

La obtención de líneas de concordancia, *collocates*, *collligates*, *bi- grams* y de *clusters* nos descubriría aspectos léxico-gramaticales del lenguaje utilizado por los miembros de la comunidad discursiva. Mediante este procedimiento pudimos detectar aquellos patrones recurrentes comunes a los diferentes textos analizados y, consecuentemente, característicos del lenguaje que representan.

Las matrices resultantes de aquellos ejemplos de co-selección léxica y gramatical nos abrieron las puertas hacia esa red semántica del conocimiento disciplinar. Partiendo de la idea de las redes sociales, y haciendo uso de *Netdraw*, tomaríamos nuestro Corpus UPV como si de una organización se tratase y en la que sus integrantes serían las diferentes unidades léxicas y las estructuras en que éstas se integran. De la misma manera en que el análisis de las redes sociales se interesa por la consistencia de las relaciones entre los actores de la organización, es decir, sus vínculos más o menos estrechos, en nuestro modelo nos interesamos por el peso de las asociaciones entre los elementos que conforman ese entramado que es el lenguaje.

Nuestra contribución en este aspecto ha consistido en diseñar un procedimiento por el que diferentes aspectos intertextuales e intratextuales de los documentos analizados pueden disponerse de tal forma que se aprecien los lazos existentes entre los diversos actores (elementos del corpus) sometidos al análisis. En este sentido, las formulaciones de Hoey (1991 y 2001) y su concepción de los conjuntos de textos como formaciones reticulares, generadas a partir de matrices y diagramas de flujo (*flow charts*), han estado presentes a la hora de plantearnos esta hipótesis como punto de partida y su constatación ha sido uno de los objetivos primordiales de la investigación.

No obstante, cabría mirar hacia atrás a nuestra hipótesis de partida, y concluir esta revisión afirmando algo sobre nuestros planteamientos iniciales: el estudio y la representación del conocimiento explícito a través del lenguaje reviste una complejidad tal que, en principio, requeriría acotarse a dominios de dimensiones manejables.

El problema fundamental reside en saber cómo formalizar lo que es realmente significativo de toda esa enorme cantidad de información que se puede obtener de un corpus. Dicho en otras palabras, se necesita disponer de unos parámetros cuantitativos para determinar aquello que se puede considerar como información significativa y relevante.

V. REFERENCIAS

Bajtín, M. (1986). *Speech Genres and Other Late Essays* (Trad. Vern W. McGee). Austin, TX.: University of Texas Press

Barabási, A. L. & Jeong, H. (2002). Evolution of the social network of scientific collaborations. *Physica A:* 311(3-4), 590-614.

Barabási, A.L. (2002). *Linked. The New Science of Networks*, Cambridge, Perseus.

Barthes, R. (1970). S/Z. Paris, Seuil.

Beaugrande, R. de & Dresssler, W. (1981 [1972]). *Introduction to text linguistics*. Austin, TX: University of Texas Press.

Borgatti, S.P., Everett, M.G. y Freeman, L.C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard: MA.

De Solla Price, DJ. (1965). Networks of scientific papers. *Science,* 149, 510-5.

Fairclough, N. (2002). Language in New Capitalism. *Discourse & Society* 13: 2, 163-166.

Ferrer i Cancho, R. & Solé, R. (2001). The Small World of Language. *Proceedings of the Royal Society of London* (B), 268, 2261-2265.

Granda de, J.I., García, F., Roig, F., Escobar, J., Gutiérrez, T. & Callol, L. (2006). Redes de coautoría y colaboración de las instituciones españolas en la producción científica sobre drogodependencias en biomedicina 1999-2004. *Trastornos Adictivos,* 8, 78-114.

Granda de, J.I., F. García, F. Roig, J. Escobar, T. Gutiérrez & L. Callol (2005). Las palabras clave como herramientas imprescindibles en las búsquedas bibliográficas. Análisis de las áreas del sistema respiratorio a través de Archivos de Bronconeumología. *Archivos de Bronconeumología,* 41, 78-83.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2004). Lexical priming and the properties of text. En Alan Partington, John Morley and Louann Haarman (Eds.), *Corpora and discourse,* 385-412.

Kretschmer H. (1994). Coauthorship networks of invisible college and institutionalized communities. *Scientometrics,* 30, 363-9.

Kristeva, J. (1966). Word, dialogue and novel. En T. Moi (Ed.), *The Kristeva Reader*. Nueva York: Columbia University Press, 1986, 34-61.

Mehler, A. (2007). Large Text Networks as an Object of Corpus Linguistic Studies. En Lüdeling, Anke & Kytö, Merja (Eds.), *Corpus Linguistics. An International Handbook,* Berlin/New York: de Gruyter.

Melin, G. & Persson, O. (1996). Studying research collaboration using coauthorships. *Scientometrics*, 36, 363-77.

Molina, L. & Muñoz, J.L. (2002). Redes de publicaciones científicas: un análisis de la estructura de coautorías. *REDES-Revista hispana para el análisis de redes sociales*, 1-3. [Documento de Internet disponible en: http://www.revista-redes.rediris.es. Consultado 12-09-2007]

Newman, M. (2001). Scientific collaboration networks. Network construction and fundamental results. *Physical Review*. [Documento de Internet disponible en: http://www.personal-mich.edu/~mejn/papers/016131.pdf. Consultado 15-05-2007].

Sanz, L. (2003). Análisis de redes sociales: o cómo representar las estructuras sociales subyacentes. *Apuntes de Ciencia y Tecnología*, 7, 21-9.

Schuemie M.J., Weeber M., Schijvenaars B.J., van Mulligen E.M., van der Eijk C.C., Jelier R., Mons B. & Kors J.A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics,* 20, 2597-2604.

Scott, M. (2004). *WordSmith Tools version 4*. Oxford: Oxford University Press.

Shah, P.K., Perez-Iratxeta, C., Bork, P. & Andrade, M.A. (2003). Information extraction from full text scientific articles: where are the keywords?: Evaluation Studies. *BMC Bioinformatics*, 4, 20.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation,* Oxford: Oxford University Press.

Solé, R. V., Corominas, B., Valverde, S. & Steels, L. (2005). *Language networks: Their structure, function and evolution*. Technical Report 05-12-042. Santa Fe Institute Working Paper.

Stubbs, M. (1996). *Text and Corpus Analysis.* Oxford: Blackwell.

Stubbs, M. (2001). *Words and Phrase.* Oxford: Blackwell.