UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

www.um.es/ijes

Corpus Linguistics, Network Analysis and Co-occurrence Matrices

KEITH STUART & ANA BOTELLA[1]
UNIVERSIDAD POLITÉCNICA DE VALENCIA

ABSTRACT
This article describes research undertaken in order to design a methodology for the reticular representation of knowledge of a specific discourse community. To achieve this goal, a representative corpus of the scientific production of the members of this discourse community (Universidad Politécnica de Valencia, UPV) was created. The article presents the practical analysis (frequency, keyword, collocation and cluster analysis) that was carried out in the initial phases of the study aimed at establishing the theoretical and practical background and framework for our matrix and network analysis of the scientific discourse of the UPV. In the methodology section, the processes that have allowed us to extract from the corpus the linguistic elements needed to develop co-occurrence matrices, as well as the computer tools used in the research, are described. From these co-occurrence matrices, semantic networks of subject and discipline knowledge were generated. Finally, based on the results obtained, we suggest that it may be viable to extract and to represent the intellectual capital of an academic institution using corpus linguistics methods in combination with the formulations of network theory.

KEYWORDS: corpus linguistics, co-occurrence matrices, semantic networks, knowledge discovery.

RESUMEN
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

[1] ***Address for correspondence***: Keith Stuart / Ana Botella. Departamento de Lingüística Aplicada. Universidad Politécnica de Valencia. 46022 Valencia; Tel.:96-652-8495. Spain. E-mail*: kstuart@idm.upv.es*

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. INTRODUCTION

This article proposes a model for the application of network analysis to the field of corpus linguistics as a method for the representation of the knowledge that is generated in our academic discourse community. The initial idea is a simple one: the words that conform a corpus are the nodes of an interrelated linguistic network. The article analyzes the discourse of science and technology by means of the study of keywords and their co-selection in research articles belonging to a corpus of 1,376 articles (a total of 6.104.323 words). All of the articles have been taken from specialist journals and have been written by our academic staff and represent the work of a unique discourse community. These articles have been published in journals that are indexed in the *Science Ctation Index* (*SCI*®).

The hypothesis which we started from in our investigation is that language, and in this case written text, is the vehicle of exchange and transmission of knowledge between the members of a discourse community. What we are dealing with here is an attempt to extract the knowledge that has been shaped in scientific articles, to analyze it and to organize it so as to be able to represent it. To achieve this, we made use of our selected corpus of journal articles and their analysis, the microscopic and macrocospic study of certain lexico-grammatical characteristics which realize networks of meaning, the knowledge that is generated in a university context. In this academic scenario, terminology extraction and analysis becomes a central issue.

According to the Firthian tradition, collocations manifest certain lexical and semantic affinities that go beyond grammatical restrictions. Sinclair (1991: 170) refers to collocation as "the occurrence of two or more words within a short space of each other in a text"; logically, this definition could be making reference to the co-selection between lexical and/or grammatical items. From the point of view of network theory, we can explain the concept in the following way: if two units *a, b* are related in terms of collocational statistics (or are simply frequent bigrams) as are units *b, c*, then there is an implicit and indirect relation between *a* and *c*, even though there has been no direct confirmation of an existing collocational relationship between *a* and *c*. We have been cautious in our assumptions in this study, using only the collocations/bigrams of those words that had been obtained as keywords and may be considered to be cohesive nodes, because they are related at least three times with other keywords (Hoey, 1991).

The article explains how we generated matrices of co-occurrences of keywords and how we visualized the co-appearance of these keywords in 23 different areas of specialized

knowledge and for the corpus in its totality. For this task, we had to use various computer programs. Wordsmith was used to extract keywords from the corpus. An initial listing of keywords was obtained by comparing our corpus (*UPV Corpus*) of English research articles with a corpus of general English (*British National Corpus*). At the same time, listings of keywords of each one of the 23 specialist areas were obtained by comparing the initial listing of keywords extracted from the UPV corpus with each of the specialist areas (*key-key words*). The matrices of keywords were made by means of a program we developed using Perl and dumped onto spreadsheets. At this point, each one of the matrices was transferred to the Ucinet program and, finally, the networks were visualized with the Netdraw utility.

A high-priority objective of the article is to show how these intratextual and intertextual networks generated from the keywords offer granular fragments of knowledge that are dispersed within, throughout and across texts, and contain a high semantic load. Advances in network theory not only provide a suitable framework of integration, but they may open new perspectives in the study of language and the organization of knowledge. Corpus linguistics in combination with network analysis may become a technique applicable to the discovery of knowledge and, in our particular case, disciplinary and subject knowledge.

II. METHOD

In the study, we have been able to discover how words used in scientific terminology dependent on a specialized field of knowledge, generally display low frequency statistics in the normal discourse of general English. These specialized terms help to define the communities that use them in the same way as these communities define their terms. The information compiled in the different stages of the research has made use of the notions of word frequency, keywords and lexico-grammatical relations, that is to say, the lexico-grammatical phenomena of collocation, semantic prosody and colligation. Similarly, basing ourselves on statistical relevance, we have evaluated the degree of interaction, the associations that take place between certain lexico-grammatical items in our research.

Besides the intratextual study realized, certain intertextual aspects have been considered that have allowed us to detect variations which are produced within the same genre. For this purpose, we have worked in the development of computer applications designed to suit our needs. We have been able to compare our tools with other existing commercial tools on the market that have similar aims, such as for example Wordsmith Tools. Both the advantages and the weaknesses of these tools as well as the results obtained after their use have been compared. These questions have been addressed by analyzing our UPV corpus in a general and global manner, as well as for each one of the specialized knowledge areas within the corpus.

Once concluded the intratextual analysis, in the following stage, an analysis was carried out that allowed us to quantify and to represent concrete aspects about variation and recursivity at the intertextual level. We started from the premise that, over and above

individual texts, there exist textual macrostructures that various texts share or is generic to them and that it is possible to access these macrostructures by means of corpus linguistic methods.

Authors such as Kristeva (1966), Barthes (1970) or Bakhtin (1986) understand intertextuality in the sense that a text is always tied to other texts or previous experiences and show prospection to future texts or wordings and statements. The intertextual acts of retrospection and prospection means that the interactive force of a text extends back to previous texts and forward to future texts. De Beaugrande and Dressler (1981) affirm that any text must fulfill the requirement of intertextuality so that it can be considered itself to be a text and that, in addition, intertextuality determines the way that the use of a certain text depends on the knowledge of other texts. For these authors, the term intertextuality refers to the dependency relation that is established between the processes of production and reception of a certain text and the knowledge that the participants in the communicative interaction already have of other previous texts related to the text in question.

Along the same lines, Fairclough (2002) defends an intertextual perspective for the analysis, for example, of pre-constructed phrases and fixed collocations.

Once delimited the framework for this phase of the study, we defined as specific objectives:

To represent the frequency of each keyword in each of the different documents that make up the areas of knowledge within the UPV Corpus

To represent the distribution of each of these keywords in the different sections that traditionally form part of the research article (*IMRD*)

To relate and to represent the interactions between terms according to their frequency rate

To compare and to represent the degree of recursivity that is produced with regards to identical language patterns of different length (clusters) in each one of the analyzed texts

The work was carried out in four successive stages that are shown in the following table:

Matrix generation: Intratextual and intertextual analysis
Matrix 1: Keyword distribution per document
Matrix 2: Keyword distribution per article sections
Matrix 3: Keyword combinations
Matrix 4: Cluster distribution (3 to 8 words) per document

Table 1. Matrix Generation

The basic scheme that was followed for each one of the matrices is as follows:

Matrix 1

	Doc 1	Doc 2	Doc 140
Word 1	Frequency		
Word 2			
Word 3			
Word *n*			

Matrix 2

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

Matrix 3

	Result	*System*	Word *3*			*Word 100*
Word 1	Frequency					
Word 2						
Word 3						
Word *n*						

Matrix 4

	Doc 1	Doc 2	Doc 140
Clusters 3 Words	Frequency		
Clusters 4 Words			
Clusters n = 8			

Table 2. Scheme for matrix generation

The matrices were generated from the lists of keywords of each area of knowledge and from keywords in the corpus in its totality. A software application that we developed ourselves in Perl was used for this and each of the matrices was transferred to a spreadsheet.

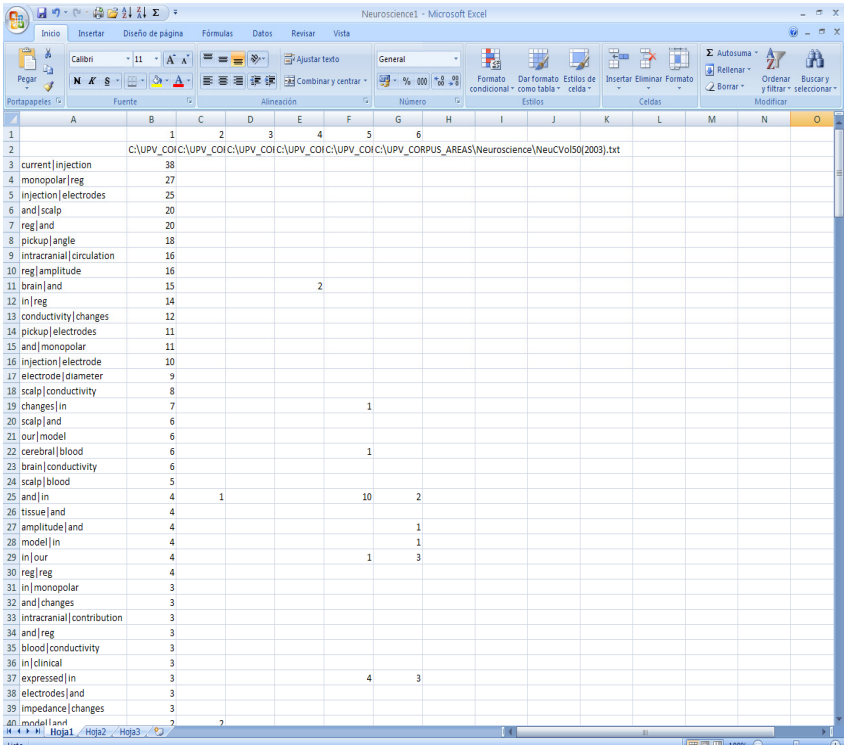
The following phase consisted in valuing and determining which computer program would be the most adequate to carry out the representation of information in reticular form from those intratextual and intertextual aspects that had been obtained in the form of

matrices. Ucinet 6 demonstrated to meet the conditions for such aims. For this reason, using the Netdraw utility of the tool, we proceeded to carry out different representations that allowed us to establish conclusions about the graphical representation of knowledge from the matrices of co-occurrences of keywords.

Ucinet is a tool for the representation of social networks. The analysis of social networks constitutes a method for evaluating informal networks by means of the representation of the relations between people, equipment, departments or even whole organizations. It studies the form in which individuals or organizations are connected and defines the position that these occupy in the network, the groups and global structure of the network, knowledge and information flows within the network and network relations which involve reciprocal influence. For a number of years, this kind of analysis has been applied to investigate ongoing collaboration between authors or institutions in scientific publications. Examples of this kind of research initiative can be found in Newman (2001), Molina and Muñoz (2002), Sanz (2003), González Alcaide et al. (2006).

III. RESULTS

In *Matrix 1* pairs of keywords from each of the documents obtained from the individual areas that make up the UPV Corpus are represented. By this method, those pairs that are specific to a single article as well as those that are repeated in more than one text can be identified. The matrix we have selected as an example corresponds to the area of Neuroscience. As it is a knowledge domain with a reduced number of articles, it is possible to visualize a screenshot in which the distribution of the items in the spreadsheet is shown.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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												

Figure 1. Screenshot Matrix 1: Bi-grams per document

The network we present below (fig.2) demonstrates how the majority of bi-grams are usually grouped around single documents in our corpus. In contrast, some of them, especially those with a lower semantic load, share intermediate positions as they are found in more than one article. The results obtained after the first stage in our analysis lead us to confirm that pairs of keywords with high semantic density tend to concentrate in individual texts, which denotes the specificity of the articles analyzed.

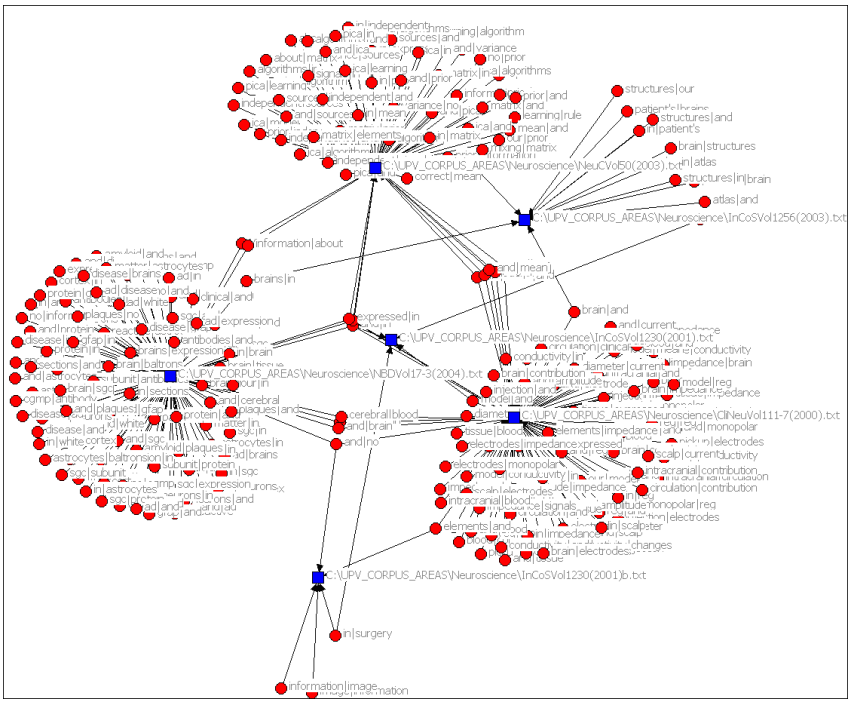


Figure 2. Network example: Bi-grams per document

The following matrix, Matrix 2, represents the distribution of keywords across the different sections in which academic research articles tend to be structured (Abstract, Introduction, Methods, Results, Discussion, Conclusion). The information it provides offers clear indications with regards to what is known as the 'aboutness' of the texts that make up the specialised knowledge subdomains or sub-corpora. When analyzing keyword lists from the individual areas in previous stages in our study, we obtained global data referring to implicit knowledge. At this stage, we have the necessary tools to interpret quantitatively how that lexical information with knowledge content is structured in the standard sections of academic articles. This issue has been addressed by various different authors, from diverse specialist areas, who base their studies on text mining to discover knowledge that is present in a large number of texts and which would be impossible to extract manually only by means of exhaustive reading. At this point, it should be emphasized that the majority of studies have conducted their analysis only by processing the Abstract section of articles. A similar analysis, based on keyword distribution in academic article sections, was developed by Shah et al. (2003). The reason for concentrating on the Abstract section responds, on the one hand, to the availability of abstracts online and, on the other, to the large amount of information that is condensed in them. Nevertheless, the Results section is the one that covers a greater quantity of information within the article, whereas Abstracts contain a greater density of information (Schuemie et al., 2004).

If we observe table 3, taken from the area of Chemistry, showing the distribution of keywords in each sections, we will discover that in the Abstract section terms like 'compound`, 'polymeric', 'immunoassay' and 'pesticides' are repeated significantly (although this section is greatly reduced in extension). In the Method and Results sections, terms that were found to be statistically relevant are, for example: 'curve', 'fig', 'observed', 'concentration/s', 'range', 'calibration', which are used to express findings after a process or model of investigation. The information we obtained by analysing the occurrence and distribution of keywords in article sections leads us to conclude that there exists a certain preference or concentration in the use of certain terms in the different sections in academic articles. As Hoey would say, article sections are lexically primed for certain words. We could, even, state that these can be grouped under categories since they tend to show common lexical and/or grammatical features.

Word	**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
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

Word	Abstract	Introduction	Methods	Results	Conclusion
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

Table 3. Distribution of 50 keywords across document sections (Chemistry)

The distribution of the terms collected in the form of matrices can be visualized using the *Netdraw* utility. When clicking each of the terms, we will see their number of links and the different categories they connect to (the nodes of the network), in this case, the different sections of an article. This procedure allows us to visualize how a term is contained in one or more article sections.

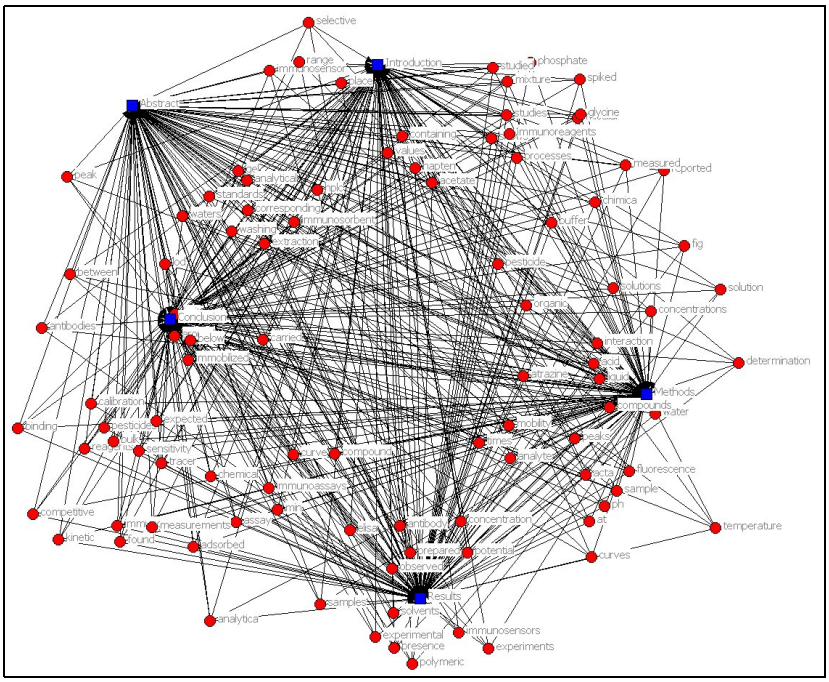


Figure 3. Network example of keywords per document section

In Matrix 3 the combinations between keywords (*bi-grams*) have been represented. The objective of this type of analysis is to determine how the same element is related to a greater or lesser extent with other relevant elements within the subject area. For this reason, the matrix was designed in both the vertical and horizontal axes including the same elements: the keywords from each area. The numbers in the cells indicate the number of combinations that take place between these pairs. The distribution of the information by this method allows the researcher to detect whether the co-occurrences are unidirectional or bidirectional, as well as the number of repetitions. For example, we could verify that the combination 'apical end' is very frequent (95 repetitions), as is also 'basal end' (87 appearances), whereas the combination 'apical bud' (3 instances) is much less common. However, at this stage, we proceeded by asking the following:

1. Are the combinations 'apical end' and 'basal end' unidirectional or bidirectional?
2. Which other terms does the term 'apical' interact/combine with ?
3. Which other combinations are found with 'basal' and 'end'?

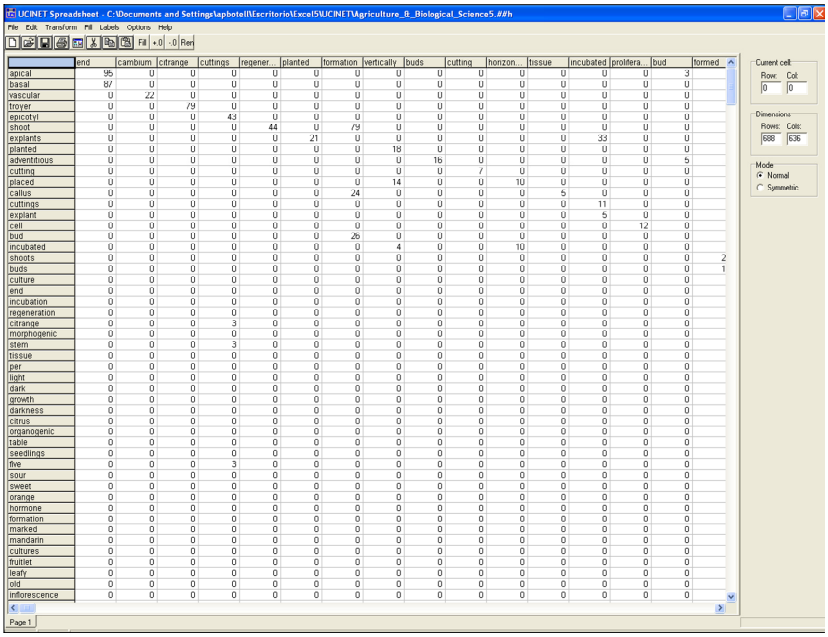


Figure 4. Matrix example of combinations between keywords (*Agriculture & Biological Sciences*)

When exploring the table, we observe that 'apical end' is only used in one direction, whereas 'basal end' is bidirectional, even though 'end basal' is less frequent (3 repetitions). In response to the second question, we can see that 'apical' also co-appears with 'bud' and 'shoot'. Moreover, we find the combinations 'adventitious bud' and 'bud formation'. We could expand the interaction or co-selection of keywords further in this way.

Likewise, when looking into the matrix for the combinations of 'end' with other terms, examples such as 'end table'and 'stylar end' are found. We discover that 'stylar' does not co-appear with other keywords. 'Basal' is combined with 'medium' (3 repetitions), with 'diet' (20 instances). When taking for our analysis a knowledge domain with a large number of texts, the matrix generated is also of great dimensions. Consequently, when trying to represent the content of this complex matrix graphically, we discover that the resulting network is a complex one-which denotes the complexity of language-in which all the existing bonds are displayed.

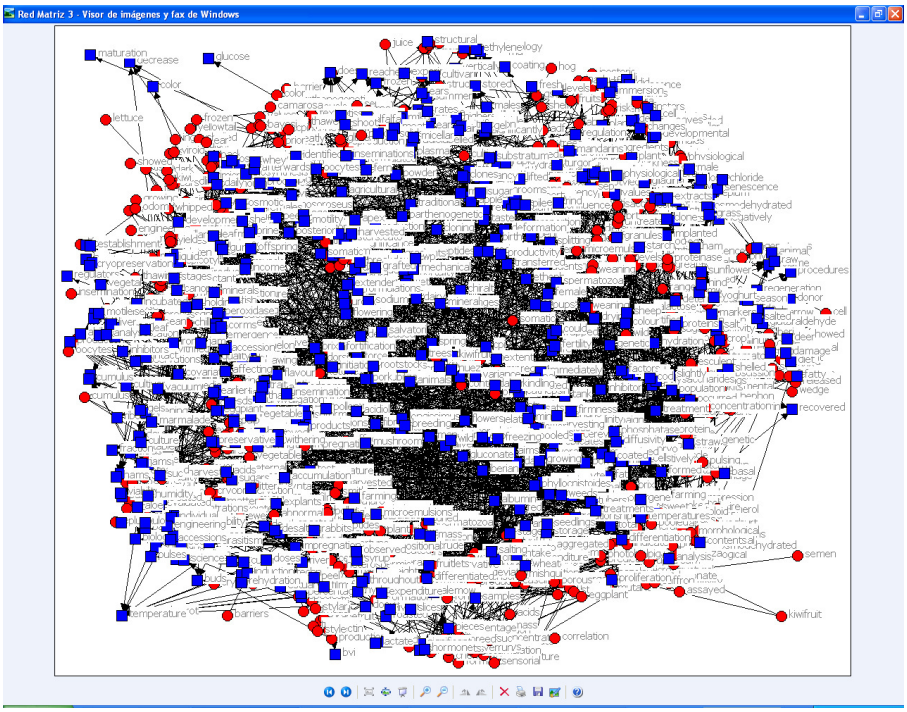


Figure 5. Network example of keyword combinations (Agriculture & Biological Sciences)

In this maze of interactions the computer tool used offers the option to apply a filter with a minimum number of appearances, so that the lines/links below the established number will be transparent, although it is also possible by selecting the 'ego' option in the tool bar to position on one of the elements and visualize solely the relationships that this participant of the network displays.

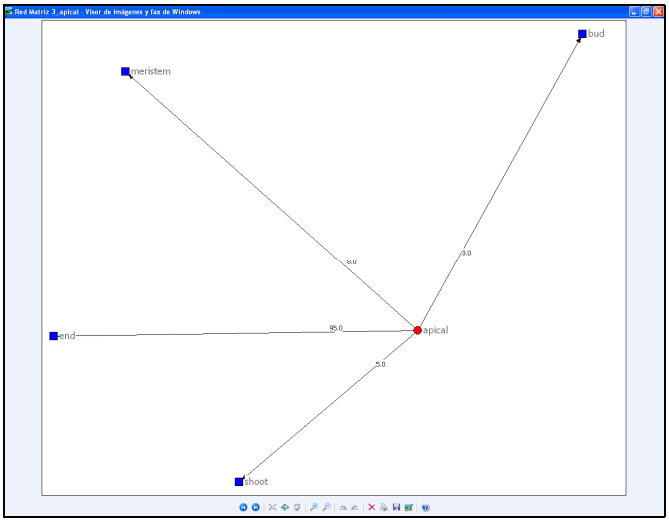


Figure 6. Network example of keyword combinations: 1 term (Agriculture & Biological Sciences)

Similarly, the utility allows us to perform multiple queries, by selecting the required elements, for example the -n most frequent keywords, and to represent their relationships. In Figure 7, the network generated from the 10 first terms in the matrix is shown.

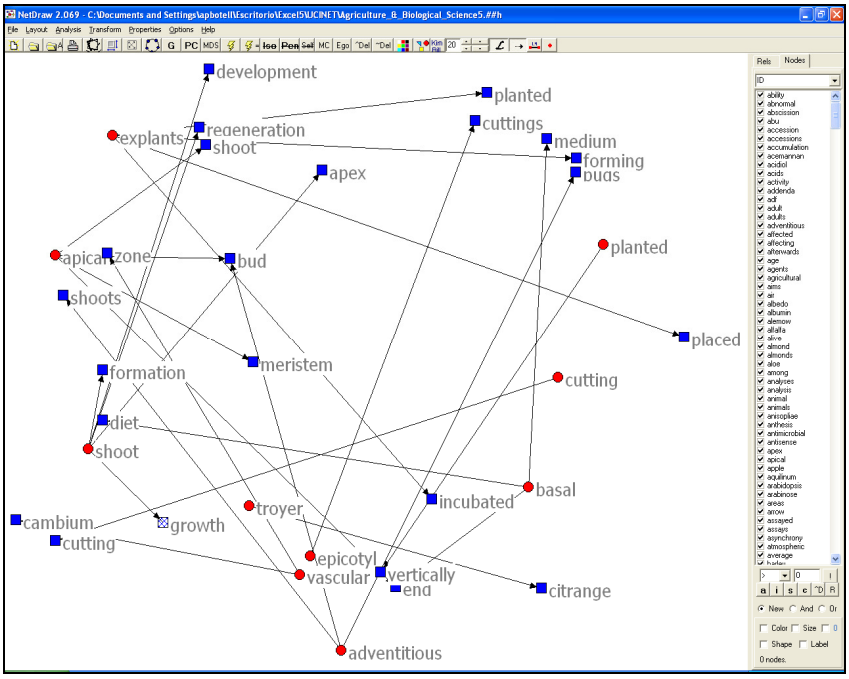


Figure 7. Network example of keyword combinations: 10 terms (Agriculture & Biological Sciences)

It is up to the researcher to decide on the selection of items to be visualized depending on the scope of her or his analysis: on the one hand, s/he might want to visualize all the relationships/bonds a term presents; thus, obtaining a conceptual dispersion, that is to say, the network will cover extensively the different concepts within the documents analysed. On the other hand, the linguist could also focus her or his study only on those combinations that are strongest, that is to say, more frequent, which therefore will have a higher conceptual and semantic density.

What we are dealing with is what could be called *social networks of language* in which the individuals or *actors* are not the members of a group, but terms, and the links are the relationships among them. Metaphorically speaking, in the same way as in social networks, we are dealing with considerations regarding the type of interactions between individuals: the number of times our participants, that is, our keywords, meet certain users in the system will imply a more or less significant/relevant relationship (in our case, conceptual and semantic density). However, the total number of participants that relate to the same actor, let alone the number of times they meet, will imply a greater complexity in the network, although its strength or consistency may be lower. With this analysis we have developed a lexical framework that has allowed us to generate maps or networks representing the explicit knowledge being produced in our academic discourse community.

The following matrix, Matrix 4, contains *clusters* or accumulations (strings ranging from 3 to 8 words) extracted from each article in the different specialized knowledge areas in the UPV Corpus, and also from the Corpus as a whole.

	ABVol84-6(1999).txt	**ABVol85-1(2000).txt**	**ABVol86-1(2000).txt**	**ABVol87-6(2001).txt**	**AE&EVol95-1(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

Table 4. Example of cluster distribution (3 words) across documents (Agriculture & Biological Sciences)

When analyzing the terms in the UPV corpus in previous stages by extracting strings of identical recurrent patterns, we could verify that, depending on the span we set, we will obtain structures with different lexical and grammatical features. In shorter sequences, like the ones shown in the table above, we detected expressions that are shared by more than one area, since they are frequent expressions in academic articles. In most cases, they are patterns that tend to be repeated in the majority of texts. The following network facilitates the visualization of this aspect:

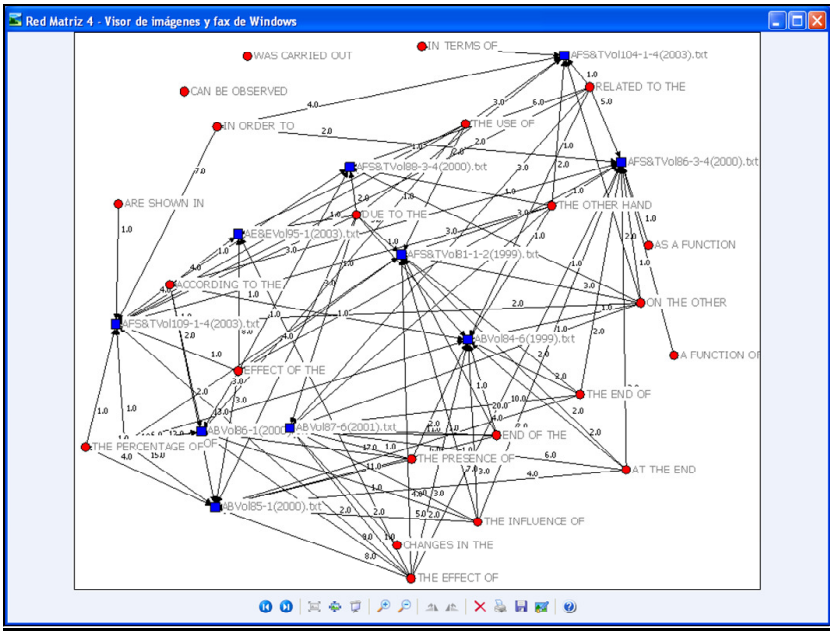


Figure 8. Network example of clusters (3 words) per document (Agriculture & Biological Sciences)

However, as strings become longer, it is observed that their semantic content is higher and, therefore, also the higher the conceptual information they convey.

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

Table 5. Example of cluster distribution (6 words) across documents (Agriculture & Biological Sciences)

The resulting network from such an analysis demonstrates that these clusters, as they contain denser and domain-specific conceptual information, are more characteristic of a limited number of articles.

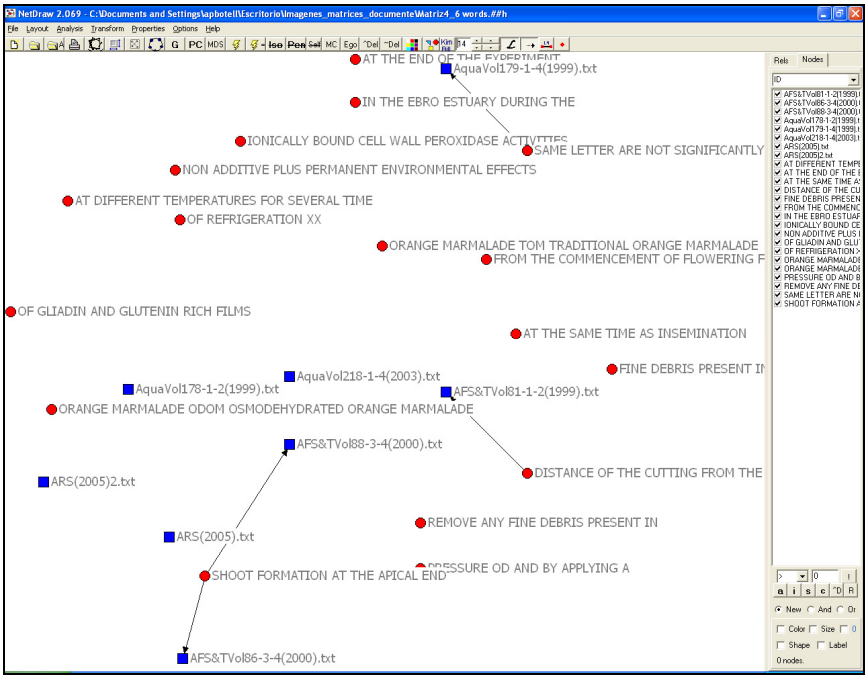


Figure 9. Network example of cluster distribution (6 words) per document (Agriculture & Biological Sciences)

We conclude our analysis of the results obtained after the methodological approach we have implemented with a statement about the complex nature of language: "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 to over long periods of time, as they adapt and change them to their needs local as part of ongoing interactions" (Solé et al., 2005: 3).

IV. CONCLUSION

Our intention has been to represent discourse as a network of meanings. In this attempt to outline a model for the generation of semantic networks basing ourselves on the idea of social networks (Barabási, 2002; Barabási & Jeong, 2002), and making use of the necessary computing tools to achieve our aim (Borgatti, 2003), we have carried out the study taking as our point of reference the principles of corpus linguistics, an empirical method that has been shown to be an adequate procedure to be able to obtain necessary information on language and knowledge.

By obtaining concordance lines, collocates, collligates, bigrams and clusters, it was possible to discover lexico-grammatical aspects of the language used by members of the discourse community being studied. As a result of this procedure, we could detect those recurrent patterns common to the different texts analyzed and, consequently, characteristic of the language that they represent.

The resulting matrices of lexical and grammatical co-selection examples have opened the doors for us to work towards a semantic network of disciplinary knowledge. Starting off from the idea of social networks, and making use of Netdraw, we analyzed our UPV Corpus as if we were dealing with an organization and whose members would be the different lexico-grammatical units and the structures into which they are integrated. In the analysis of social networks, one is interested in the consistency of the relations between the actors of the organization; that is to say, their stronger or weaker ties. In a similar manner, in our model we were interested in the weight of the associations between the linguistic elements that conform the language network.

Our contribution in this aspect has consisted of designing a procedure by which different intertextual and intratextual aspects of the analyzed documents can be obtained in such a form that one can appreciate the existing bonds between the diverse actors (elements of the corpus) that have been submitted to analysis. In this sense, the ideas of Hoey (1991, 2001) and his conception of sets of texts as network formations have been present when formulating the hypothesis that language is recursive and forms a network of meanings that carry the semantic content of texts. The establishment and verification of a relationship between these networks of meaning and knowledge has been one of the principal objectives of the investigation.

However, at this stage, we should look back at our point of departure, our initial hypothesis and conclude this article affirming that the study and the representation of explicit knowledge through language because of its complexity needs to be limited to

specialist knowledge areas of manageable dimensions. The fundamental problem resides in knowing how to formalize what is really significant out of the enormous amount of information that can be obtained from a corpus. Stated in other words, there is a need for quantitative parameters to determine what should be considered significant and relevant information.

V. REFERENCES

Bakhtin, 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. [http://www.revista-redes.rediris.es Accessed on 12-09-2007]

Newman, M. (2001). Scientific collaboration networks. Network construction and fundamental results. *Physical Review*. [http://www.personal-mich.edu/~mejn/papers/016131.pdf Accessed on 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.

