ARTÍCULOS

Círculo de Lingüística Aplicada a la Comunicación
ISSN: 1576-4737

http://dx.doi.org/10.5209/clac.68972

EDICIONES COMPLUTENSE

Progresión temática y cohesión textual a través de grafos de coocurrencias

Antonio Rifón[1]

Recibido: 12 de febrero de 2019 / Aceptado: 2 de marzo de 2020

Resumen. Este artículo estudia la posibilidad de aplicar la teoría de grafos o redes al análisis de la progresión temática y la cohesión textual. Para ello, se analizan seis textos de divulgación científica: tres sobre el mismo tema, el aceite de palma, otro sobre el aceite de oliva y dos de control; uno creado por la unión de dos textos de temas diferentes y el otro por párrafos con distintas temáticas seleccionados aleatoriamente. Se ha creado, de cada texto, un grafo no dirigido y pesado de coocurrencias con ventana 5-gram. En primer lugar, se han analizado las medidas globales del grafo para conocer su topología; en segundo lugar, se ha empleado el grado de intermediación de los lemas para conocer los temas de cada texto y se ha estudiado como estos evolucionan; en tercer lugar, a través de la modularidad del grafo, se ha analizado en qué párrafos aparecen los diferentes temas y cómo ha evolucionado.

Palabras clave: Progresión temática; cohesión léxica; teoría de grafos; medidas de centralidad.

[en] Thematic Progression and Textual Cohesion trough Graphs of Co-occurrences

Abstract. This paper studies the possibility of applying the theory of graphs or networks to the analysis of thematic progression and textual cohesion. To this end, six scientific divulgation texts are analyzed: three on the same subject, palm oil, another on olive oil and two control texts, one created by the union of two texts of different subjects and the other generated randomly by paragraphs on different subjects. An undirected and weighted graph of co-occurrences with a 5-gram window has been generated from each one. First, we have analyzed the global measures of the graph to know its topology; secondly, we use the degree of betweenness to know the topics of each text and we study how these evolve; thirdly, through the modularity of the graph, it has been analyzed in which paragraphs the different topics appear and how they have evolved.

Keywords: Thematic Progression; lexical cohesion; Graph Theory; centrality mesaures.

Cómo citar: Rifón, A. (2020). Progresión temática y cohesión textual a través de grafos de coocurrencias. *Círculo de Lingüística Aplicada a la Comunicación* 82, 193-208, http://dx.doi.org/10.5209/ clac.68972

1. Introducción

Un grafo o red es un conjunto de puntos, a los que llamaremos nodos, unidos por líneas, a las que llamaremos aristas (Newman, 2010, p. 1), que representan las relaciones que existen entre los nodos. Todo texto puede ser representado como un grafo, con las palabras como nodos y las relaciones entre ellas como aristas (Mihalcea & Radev, 2011). En este trabajo analizamos cómo pueden ser empleados los grafos, cuyos nodos son las palabras y las aristas las relaciones de coocurrencia, para el estudio de la progresión temática textual; además, hacemos un primer acercamiento al potencial de dichos grafos para el estudio de la cohesión léxica.

Un texto no es solo una lista de palabras colocadas secuencialmente, sino que tiene pretensión de unidad de sentido; para construir esa unidad, se establecen distintas relaciones entre los elementos del texto, nuestra atención se centra ahora en la repetición y las colocaciones que contribuyen de una manera importante a la cohesión léxica (Halliday & Hasan, 1976).

El concepto de colocación necesita alguna aclaración pues es un término un tanto vago y que, por lo tanto, ha sido interpretado de diferentes maneras incluso dentro de la propia Lingüística Sistémico Funcional (Tanskanen,

1 Universidade de Vigo (España). Correo electrónico: arifon@uvigo.es

2006; Ding, 2018). En este estudio el término colocación se refiere a la coocurrencia de dos lexemas dentro de una ventana previamente establecida, es una interpretación semejante a la dada por Stubbs (2001, p. 305) para quien la colocación es "a purely lexical and nondirectional relation: it is a node–collocate pair which occurs at least once in a corpus".

Si convertimos el texto en un grafo de coocurrencias al que se le aplican los métodos y herramientas de la teoría de grafos, podremos analizar, de forma cuantitativa, no solo la repetición de palabras, sino también sus relaciones lo que nos proporcionará un estudio, también, de sus colocaciones. Todo ello, nos permitirá conocer los temas fundamentales del texto, cómo han evolucionado esos temas, tomando el texto como un todo, qué temas priman en cada parte del texto y cómo varían en el proceso de escritura.

Nuestra tarea está muy relacionada con la extracción de temas (*topics)* y palabras clave (*keywords*), entendidas estas como términos que representan el contenido fundamental del texto (Ercan & Cicekli, 2007; Grineva et al., 2009), en el procesamiento del lenguaje natural y, en este ámbito, hay ya numerosos trabajos que emplean de alguna forma el análisis de grafos para dicha extracción (Mihalcea & Radev, 2011, pp. 154-162; Lahiri, 2013; Zhou et al., 2013; Bougouin et al., 2013; Boudin, 2018; Garg & Kumar, 2018), pero también con otras áreas como la calidad textual y el análisis estilístico por medio de grafos (Antiqueira et al., 2005, 2007; Amancio, 2015). De los primeros tomamos la idea de que los grafos son una herramienta adecuada para analizar la temática de los textos; pero, a diferencia de la mayoría de ellos, nuestro objetivo no es la indexación de textos, sino el estudio de la organización de textos particulares. El análisis de cómo el autor organizó temáticamente el texto, de cómo lo cohesionó, está muy en consonancia con la propuesta de los segundos.

Nuestro método e ideas también están muy relacionadas con el estudio que hace Paranyushkin (2011) de lo que él llama "pathways of meaning circulation" y que hemos traducido por circulación de significados. Su intención es rastrear las huellas, "pathways", que generan significado en la construcción del texto a través del establecimiento de grupos de conceptos por su grado de relación y la identificación de aquellos que más peso tienen en la construcción del significado textual.

No podemos olvidar, dentro de los antecedentes del trabajo, otras propuestas de patrones de cohesión como las cadenas léxicas (Morris & Hirst, 1991; Hirst et al., 1998; Barzilay & Elhadad, 1999), las cadenas nominativas o nominales (Bernárdez, 1982; Calsamiglia Blancafort & Tusón Valls, 2004; Barranco Flores, 2015) ni otras perspectivas y enfoques del análisis de la cohesión como el análisis semántico latente (Landauer & Dumais, 1997; Venegas, 2003; Landauer, 2007) y otros modelos probabilísticos (Steyvers & Griffiths, 2007).

Nuestro objetivo es, retomando las ideas de estos investigadores, aportar un método por el que se pueda estudiar, de la forma más automática y cuantitativa posible, la temática del texto, la progresión de esta, tanto del texto tomado como un producto acabado como del texto como producto en construcción temporal, y las posibilidades de estas medidas para la detección del grado de cohesión léxica por repetición y colocación léxica.

En primer lugar, describiremos los procedimientos empleados para construir el grafo de un texto (2.1.) y los textos que emplearemos como muestra (2.2.). En segundo lugar, estudiaremos los resultados (3); primero, las características generales de los grafos obtenidos (3.1.), es decir, sus medidas globales, después, la progresión temática del texto como un todo (3.2.), seguiremos con el análisis de la progresión temática del texto en el proceso de escritura (3.3.) y, finalizamos, dando una visión global con el análisis visual de los grafos (3.3.).

2. Métodos y materiales

2.1. La construcción de los grafos

Para este estudio nos propusimos crear un procedimiento que fuese lo más simple posible, que dependiese los menos posible de la lengua del texto, que no emplease información procedente de bases semánticas u ontológicas externas y que permitiese trabajar con categorías gramaticales.

El primer paso para la construcción de los grafos es lematizar el texto; necesitamos eliminar los morfemas gramaticales de las palabras para conseguir un texto compuesto por los temas o lemas. Para ello empleamos Freeling (Carreras et al., 2004; Padró & Stanilovsky, 2012), una librería de software libre que nos permite lematizar con gran fiabilidad, además de poder ser aplicada a muchas otras lenguas desde el inglés hasta el ruso y que podría ser ampliada a nuevas lenguas en caso de necesidad. De todas sus posibilidades nos interesa el "PoS tagging" que genera un fichero con los datos que necesitamos: la palabra, el lema y la categoría.

A continuación se exportan los datos de la lematización a una base mysql y, automáticamente, se asigna a cada palabra el número de oración y párrafo en el que aparecen. Asignar la oración y el párrafo nos permitirá establecer en el grafo la linea temporal de construcción del texto. Se obtiene, así, una base de datos con la palabra gramatical, su lema, su categoría gramatical y la oración y el párrafo. Tenemos, pues, dos bases cada una con un texto diferente: la primera, contiene un texto compuesto por las palabras gramaticales, que sería el mismo que el original y, la segunda, un texto formado por los lemas de esas palabras, que será en el que centre nuestro análisis.

Del texto lematizado se pueden escoger las categorías de palabra que queremos que formen el grafo; para nuestro caso, escogeremos solo los sustantivos y los verbos ya que son los elementos del texto que llevan la mayor carga conceptual y que es, en estos momentos, el aspecto que nos interesa. Así que eliminamos de la base el resto

de categorías y las *stopwords* para obtener lo que llamamos el texto limpio que es sobre el que generaremos el grafo.

Como ya se ha indicado, en un grafo de un texto, las palabras, en nuestro caso, los lemas, se constituyen como nodos, también llamados en algunas disciplinas vértices, y las relaciones entre ellas o ellos, como aristas, llamadas en algún ámbito ejes. Así que, en primer lugar, tenemos que crear los nodos; para ello, construimos una base con los diferentes lemas existentes en el texto limpio y la oración y el párrafo en el que tienen su primera y última aparición. Tenemos ya los nodos de nuestro grafo.

La generación de las aristas es más compleja y para ella se han de tomar y justificar algunas decisiones. En primer lugar, hay que determinar el tipo de relación entre palabras; en segundo lugar, habrá que delimitar si se establecen fronteras o no en las relaciones, es decir, si las relaciones se circunscriben a la oración, al párrafo o a otra característica del texto; en tercer lugar, el tipo de grafo que se quiere establecer; y, por último, la extensión y el grado de relación entre palabras.

Sobre la relación entre palabras –primera decisión– tendremos en cuenta la coocurrencia que, como ya se ha indicado, permite estudiar la repetición y las colocaciones. Halliday y Hasan (1976, p. 318) al tratar la cohesión diferenciaron dos tipos, la gramatical y la léxica; dentro de esta última, entre otros aspectos, distinguen la repetición ("reiteration") y la colocación. La repetición es "the repetition of a lexical item, or the occurrence of a synonym of some kind" (1976, p. 318), mientras que la colocación se da cuando una palabra "tends to occur in the same lexical enviroment" (1976, p. 319). Ahora bien, hay que reconocer que ambos conceptos no están exentos de problemas y los distintos autores que los han tratado les han dado diferente importancia y han hecho de ellos distintas subclasificaciones (Morris et al., 2003; Tanskanen, 2006, p. 48).

Para el caso que nos ocupa, como ya hemos indicado en la introducción, una concepción simple y abierta nos puede servir, así que el grafo que construiremos tendrá en cuenta la repetición del lema y sus colocaciones. En cuanto a la repetición, atenderemos a la de la misma palabra y, en cuanto a la colocación, a la simple coocurrencia textual. Si Halliday y Hassan (1976, pp. 284-288) ponían como ejemplos de colocación palabras que mantenían algún tipo de relación léxica (*boy-girl*, *north-south, bee-honey*) y Ding (2018, p. 87) considera las colocaciones textuales como "multi-word expressions" y distingue dos tipos, las "strong collocations (e.g. for example)" y las "habitual collocations (e.g. as recently as 1987)", nosotros no restringiremos las colocaciones ni semántica ni frecuentativamente; por ejemplo, en uno de los textos anlizados, *aceite* y *palma* coocurren en el texto once veces frente a *cobertura* y *bolllería* que lo hacen solo una, pero ambas son consideradas colocaciones textuales, eso sí, parece claro que habrá que buscar una fórmula que permita dar mayor importancia a coocurrencias repetitivas que a aquellas esporádicas.

Sobre la cuestión de las fronteras –segunda decisión– hay que tener presente que un texto está compuesto por unidades más allá de la palabra, como oraciones y párrafos. El párrafo es un elemento fundamental del texto ya que, como indica Cassany (1995, pp. 42-43), se compone de "un conjunto de frases relacionadas que desarrollan un único tema", idea que, expresada de una u otra manera, aparece en numerosos autores (Brown & Yule, 1983; Dijk & Kintsch, 1983; García Berrio & Albaladejo Mayordomo, 1983; Martínez Caro, 2014), y que, siguiendo a (Siepmann & Siepmann-Gallagher-Hannay-Mackenzie, 2008, p. 65), "make a contribution to the ongoing argument" posibilitando y guardando la coherencia entre el párrafo anterior y el siguiente. Por eso restringiremos nuestra relación de coocurrencia al párrafo; así, las palabras solo se relacionarán con las palabras de su propio párrafo, de manera que la última palabra de un párrafo no se considerará relacionada con la primera del siguiente párrafo.

Esta restricción puede ser variada dependiendo del texto, podría restringirse a la oración, a la sección, al capítulo, etc.; en nuestro caso, dado que los textos son de poca extensión, el párrafo cobra todavía más importancia pues es la mayor unidad discursiva que aporta estructura (Cassany, 1995, p. 43) y es, por tanto, la unidad que se ha de tener en cuenta.

Sobre el tipo de grafo –tercera decisión– podemos señalar que, si consideramos que la relación de coocurrencia es de doble dirección, es una relación simétrica en la que la relación entre dos palabras contiguas van tanto de la palabra que la precede a la que la sigue como de la que la sigue a la que la precede, entonces, el grafo obtenido será un grafo no dirigido.

Por último –cuarta decisión– parece claro que, en líneas generales, la relación de la palabra es mayor con sus palabras más cercanas que con sus palabras más alejadas y que, en muchos párrafos, decir que hay una relación de coocurrencia entre la primera y la última palabra sería muy arriesgado; por ello, restringiremos la relación a las dos palabras que la preceden y a las dos que la siguen, sería una ventana de 5-gram. También hemos dado mayor peso a las relaciones de contigüidad que a la simple coocurrencia, de manera que se le ha dado un peso de 2 a la relación de contigüidad y de 1 a la de coocurrencia no contigua.

Para marcar esta diferencia de peso, se establece primero una arista que relaciona una palabra con sus contiguas; así, si tenemos un texto con las palabras p_{-2} p_{-1} p_0 p_1 p_2, siendo p_0 la palabra de la que vamos a establecer las relaciones, se crean primero dos aristas, $p_{-1} \rightarrow p_0$ y $p_0 \rightarrow p_1$ con peso 2 ambas y, después, otras aristas, $p_{-2} \rightarrow p_0$, $p_{-1} \rightarrow p_0$, todas con peso 1; quedarían dos aristas con peso 1 y dos con peso 2, las que relacionan a p_0 con sus palabras contiguas p_{-1} y p_1. Queda claro que, cada vez, que se da una colocación se sumarían los

pesos, de manera que las aristas pueden tener una gran diferencia de peso, cuanto más veces coocurran dos palabras mayor peso tendrá la arista que las relacionan.

Realizadas todas estas operaciones se obtiene un grafo de coocurrencias de lemas con ventana 5-gram no dirigido pesado con las relaciones restringidas al párrafo. Una vez obtenido el grafo, solo hay que cargarlo en el programa para análisis de grafos *Gephi* (Bastian et al., 2009) y proceder a su análisis.

2.2. Los textos analizados

Los objetivos de este estudio no se centran ni en el procesamiento de una gran cantidad de textos, ni en el estudio de unos textos particulares, sino que, lo que se pretende hacer es mostrar las posibilidades del análisis de la progresión temática textual por medio de grafos; por ello, la elección de los textos no se ha centrado en su temática o su importancia, sino en algunas características estructurales ya que son textos que sirven simplemente de muestra.

Teniendo en cuenta esto, hemos escogido tres textos de divulgación científica sobre el ahora controvertido aceite de palma extraídos de la red y que, para abreviar, denominaremos *ABC* (*¿Es realmente malo el aceite de palma?*, s. f.), *Comidista* (*¿Por qué es malo el aceite de palma?* | *El Comidista EL PAÍS*, s. f.), *País* (*¿Por qué Alcampo quiere retirar el aceite de palma de sus productos, si no es tóxico ni venenoso?* | *BuenaVida* | *EL PAÍS*, s. f.). A estos textos sobre el aceite de palma, se unen tres textos de control que nos permitirán hacer comparaciones, uno con una temática diferente, el aceite de oliva, al que llamaremos *Oliva* (Pérez, 2015), otro, *PaísOlor,* en el que hemos unido dos textos, uno de los anteriores sobre el aceite de palma, *País*, y otro de otra temática diferente, los olores (*¿Malos olores?*, 2017); para acabar, hemos creado un último texto, *Retales*, a partir de párrafos escogidos aleatoriamente de internet sobre diferentes temas, en este texto, cada párrafo es sobre un tema diferente y los temas son muy distantes entre sí; hay párrafos sobre volcanes, números reales, caballos, sartenes, bolígrafos, etc.

Hemos escogido textos de divulgación científica pues son textos expostivos-argumentativos en los que se ha hecho una reformulación y recontextualización (Calsamiglia & Van Dijk, 2004) de textos técnico-científicos mediante diversas estrategias discursivas (Alcíbar Cuello, 2004; Mapelli, 2006) que suponen que la estructura y las estrategias rígidas de estos se relajen (Mapelli, 2006) manteniendo aun así los conceptos nominales y verbales como elementos esenciales para llegar comunicar los complejos conceptos técnico-científicos.

Son textos breves lo que hace que se elimine complejidad y podamos centrarnos en el método y evitar las largas explicaciones sobre aspectos concretos de los textos que sí son necesarias para el análisis de textos largos. Se han seleccionado, además, tres con la misma temática, el aceite de palma, uno con una temática relacionada, el aceite de oliva, otro con una totalmente diferente, el olor, y otro hecho de retales para poder hacer comparaciones entre diferentes grados de similitud temática. Para finalizar, teniendo en cuenta nuestros objetivos, hay que advertir que no se hará un estudio profundo de cada uno de los textos, sino que iremos escogiéndolos según el caso para mostrar los diferentes aspectos del análisis.

3. Resultados y discusión

Esta sección está dividida en cuatro subsecciones; en la primera, se analizan las características generales de los textos y las principales medidas globales de sus grafos; en la segunda, se estudian los temas y su progresión tomando el texto como un todo; en la tercera, se atiende a la progresión de los temas a lo largo de los párrafos; y, para finalizar, se hace una breve explicación del grafo en su totalidad.

3.1. Características generales de los textos y grafos

En la Tabla 1 se muestran las características principales de los textos: número absoluto de palabras (*types*), número de palabras diferentes (*tokens*), número de lemas, número de oraciones y número de párrafos.

	Palabras (Types)	**Palabras (Tokens)**	**Lemas**	**Oraciones**	**Párrafos**
ABC	833	344	301	30	11
Comidista	1812	635	536	55	25
País	1673	596	516	53	15
Oliva	1147	378	315	39	22
PaisOlor	3169	1069	883	139	32
Retales	1805	676	595	51	12

Tabla 1: Características numéricas principales de los textos analizados.

En la Tabla 2, en la página siguiente, se muestran las medidas globales de los grafos de cada texto, hay que recordar que los grafos se componen de las relaciones de coocurrencia solo de sustantivos y verbos, relaciones que han sido ponderadas o pesadas.

	ABC	**Comidista**	**País**	**Oliva**	**PaisOlor**	**Retales**
Nodos	175	325	324	157	557	374
Aristas	449	892	907	463	1677	970
Grado medio	5,131	5,489	5,599	5,898	6,022	5,187
Grado medio con pesos	8,446	9,12	9,099	11,172	9,601	8,561
Diámetro	10	11	11	7	9	11
Densidad	0,029	0,017	0,017	0,038	0,011	0,014
Modularidad	0,612	0,623	0,614	0,538	0,611	0,743
Componentes conexos	1	2	1	1	2	1
Coeficiente de clustering	0,512	0,5	0,469	0,535	0,458	0,504
Longitud media de camino	3,88	3,97	3,782	3,139	3,919	4,813

Tabla 2: Medidas globales de los grafos de cada texto.

La primera medida que llama la atención es la de componentes conexos. Tener un único componente conexo quiere decir que hay un camino para ir de cualquier nodo a cualquier otro de la red; si hay dos o más, quiere decir que, en la red, hay grupos o subgrupos de nodos desconectados. Teniendo en cuenta la forma en la que se han establecido las conexiones de las redes, lo normal es que solo haya un componente; si hay más de uno, entonces existen párrafos en los que ninguna de sus palabras (sustantivo o verbo en este caso) aparece en otros párrafos.

Todos los grafos tienen un único componente, como era de esperar, excepto *Comidista* y *PaisOlor* que tienen, inesperadamente, dos. En *Comidista* hay un breve párrafo final apelando al lector "ahora el que decide eres tú" cuyo verbo, *decidir*, no aparece más veces y, en *PaisOlor*, ninguna palabra del título del apartado "ventosidades caninas y felinas" se repite en el texto y, por tanto, ambos crean un componente no conexo del grafo.

Se puede ver también en la Tabla 2 que la densidad de los grafos es muy pequeña, es decir, se mantienen pocas conexiones entre palabras; si todas las palabras estuviesen conectadas con todas las demás, la densidad sería 1. Vemos, además, que la longitud del camino medio (Newman, 2010, pp. 136-140; Solé et al., 2010), es decir, la media de los caminos más cortos entre los nodos es pequeña con respecto al diámetro del grafo y que tienen un coeficiente de clustering bastante alto (Newman, 2010, pp. 262-265; Mihalcea & Radev, 2011, pp. 62-63). Estos datos apuntan a la existencia de una estructura de mundo pequeño (*small world*) (Watts & Strogatz, 1998; Ferrer I Cancho & Solé, 2001) en el que los nodos no están conectados con muchos otros nodos, pero que los vecinos de un nodo suelen ser vecinos entre sí; esta característica nos hace pensar que la distribución de las conexiones tiene que presentar cierto grado de asimetría.

Podemos sospechar que existe un gran número de palabras que se conectan a otras pocas palabras y que hay unas pocas palabras que se conectan a muchas otras palabras; las primeras tendrán poco grado de conexión y las segundas, un alto grado de conexión. Para ello analizaremos la asimetría y la curtosis del grado de los nodos de cada texto (Tabla 3):

	ABC	**Comidista**	**País**	**Oliva**	**PaísOlor**	**Retales**
Asimetría	5.502	6.124	7.333	5.646	6.368	3.305
Curtosis	33.524	44.871	65.083	36.503	52.297	12.081

Tabla 3: Valores de asimetría y curtosis de la distribución de los grados de los nodos

Vemos que los cinco primeros textos presentan asimetría a la derecha, la cola de la media hacia la derecha es más larga, es decir, presentan una gran cantidad de palabras con un grado pequeño de conexión, y también un alto grado de curtosis que muestra un alto grado de apuntamiento de la distribución frente a la normal. El único texto que se sale de estos parámetros es el de *Retales*, esto se debe a que la repetición léxica es pequeña, recordemos que estaba formado por párrafos de distintos temas, y, por tanto, las palabras no ganan muchas conexiones más allá del párrafo en el que aparecen por primera vez.

El mantenimiento en *Retales* de las medidas de la longitud del camino medio y del coeficiente de clustering semejante al resto de textos se explica porque cada párrafo se comporta como un pequeño texto, de manera que tenemos pequeños textos que se comportan de forma semejante al resto de textos y que, unidos, mantienen esas características; fijémonos que lo mismo ocurre en *PaísOlor*, en el que unimos dos

textos que de forma independiente van a tener una longitud del camino medio pequeña y un coeficiente de clustering alto, cuando los unimos, ninguna de estas características se ve afectada.

El hecho de que tengamos unos pocos nodos muy conectados con otros, pero que la densidad media sea baja, y que los vecinos de un nodo sean vecinos entre sí, facilita dos características importantes, en este momento: la primera, que sea más fácil establecer grupos de nodos caracterizados por tener conexiones entre sí más densas que la media, es decir, que podamos crear comunidades de nodos, la facilidad para hacer esto la marca la modularidad (Lambiotte et al., 2008) que, como vemos en nuestros casos, es bastante alta; la segunda, que habrá nodos que de una manera más clara conectarán estas comunidades entre sí, este grado de conexión lo marca la intermediación.

Nos centraremos ahora, principalmente, en estas dos medidas, modularidad e intermediación, para ver si es posible conocer la temática del texto de forma global, su progresión globalmente y la progresión temática del texto como objeto en crecimiento párrafo a párrafo.

3.2. Progresión temática global

En este apartado atenderemos a la medida de intermediación para analizar el texto como un todo. En primer lugar veremos cuáles son las palabras con mayor grado de intermediación (*betweenness*) (Brandes, 2001; Newman, 2010, pp. 185-193) y, en segundo lugar, como ha ido evolucionando su grado a lo largo de los párrafos del texto.

La intermediación nos indica el porcentaje de caminos más cortos entre nodos que pasan por ese nodo; cuantos más caminos cortos entre palabras pasen por una palabra está tendrá más grado de intermediación. Esta medida nos ayuda a detectar aquellas palabras que unen otras palabras no adyacentes, es decir, aquellas que sirven de unión para que los significados de las otras estén relacionados cuando no son vecinas; si eliminásemos estas palabras, las otras y sus significados quedarían aislados o mucho menos conectados. Si se considera que en un texto además del tema central hay subtemas relacionados con el central, estas palabras son las que facilitan esa unión, no unen palabras en grupos, sino que unen grupos de palabras; para ello son fundamentales los dos aspectos de la cohesión léxica: la repetición y la colocación.

Para que una palabra tenga una alta intermediación es necesario que se repita, pero no solo es un problema de cantidad, sino también de colocación; es necesario que se repita en aquellos puntos estratégicos del texto que sirvan para cohesionarlo y, además, es necesario que se coloque con las palabras estratégicas dentro de cada grupo. Así pues, la intermediación nos indica aquellas palabras que sirven para dar cohesión al texto, que facilitan la circulación de significados textuales que, en principio, podrían estar desconectados y son un indicio del tema o temas centrales del texto.

La intermediación se ha empleado también como una medida para seleccionar las palabras clave del texto; Boudin (2013) prueba, para la selección de palabras clave, diferentes medidas de centralidad en tres corpus y, aunque la intermediación no es ni la menos, ni la más adecuada en ninguno de los tres corpus de forma aislada, es la medida que sale mejor parada tomados los corpus en conjunto. En este momento no proponemos nuestra selección de palabras por su intermediación como las palabras clave del texto, sino como aquellas que permiten la circulación de significados (Paranyushkin, 2011), aunque, como se verá, también se podrían tomar como una buena aproximación a las palabras clave.

Para el análisis hemos escogido solo las seis primeras palabras para simplificar la explicación y eliminar la complejidad que supondría escoger muchas palabras, porque, a partir de la sexta, el grado de intermediación baja mucho, y, también, porque los modificadores de clase *palma* y *oliva* de *aceite de palma* y *aceite de oliva* han sido analizados separados de su núcleo, de manera que salvamos un poco esas construcciones cercanas a los compuestos para quedarnos con seis palabras si los consideramos separados o con cinco si los consideramos unidos (Tabla 4).

	Intermediación					
Textos	**Palabras**					
ABC	Aceite	Palma	Grasa	Producto	Experto	Industria
	0,3547	0,2275	0,2093	0,1833	0,1504	0,0818
Comidista	Aceite	Alimento	Palma	Grasa	Ácido	Ingrediente
	0,2286	0,1837	0,1652	0,1396	0,1240	0,9000
País	Aceite	Palma	Ácido	Persona	Alimentación	Colesterol
	0,3080	0,2779	0,1370	0,1017	0,0839	0,0659
Oliva	Aceite	Oliva	Alimento	Ayudar	Hígado	Colesterol
	0,3194	0,2108	0,1729	0,1009	0,0834	0,0748
PaísOlor	Aceite	Olor	Palma	Ácido	Persona	Encontrar
	0,1822	0,1696	0,1688	0,0862	0,0706	0,0507
Retales	Permitir	Número	Agua	Filtro	Radio	Adecuar
	0,1851	0,1842	0,1292	0,1266	0,1244	0,1074

Tabla 4: Grado de intermediación de las seis palabras de cada texto con el grado más alto

Con estas seis palabras se puede ver, creo que diáfanamente, cuál es el principal tema de los cuatro primeros textos y también intuir que los dos primeros –*ABC* y *Comidista*– se centran en el uso del aceite de palma como producto e ingrediente en la industria atendiendo a sus grasas y ácidos, mientras que –*País* y *Oliva*– parecen centrarse más en los efectos de los aceites en la salud, como parece indicar la aparición de *hígado* y *colesterol*, siendo en el texto *Oliva* para *ayudar* a la salud. Es indudable que, si escogiésemos un número mayor de palabras, podríamos definir mejor los temas, pero volveremos sobre ello en el siguiente apartado.

En los dos últimos textos, los textos de control, tenemos, por un lado, que, en el texto *PaísOlor*, irrumpen nuevas palabras con respecto al texto *País* y, por otro, que los temas en *Retales* son mucho más difíciles de determinar. Para analizar estos casos atenderemos a la evolución del grado de intermediación.

Para analizar la evolución, hemos calculado el grado de intermediación de las palabras según avanza la escritura, es decir, hemos tomado las palabras del primer párrafo y calculado su intermediación, a ese primer párrafo le hemos sumado las del segundo y calculado la intermediación del conjunto y así sucesivamente hasta llegar a sumar el último párrafo, es decir, hasta tener el texto completo, cuyas seis palabras con el grado más alto han de coincidir con las de la Tabla 4. Así pues, no estamos calculando el grado para cada párrafo concreto, sino el grado que se da en el proceso de escritura en el que influye no solo el último párrafo escrito, sino, también, todo lo escrito anteriormente.

Comparemos ahora la evolución de la intermediación de las seis palabras de la Tabla 4 en los textos de control –*PaisOlor* (Figura 1) y *Retales* (Figura 2)– para, después, comparar el gráfico de estos textos con el de alguno de los otros.

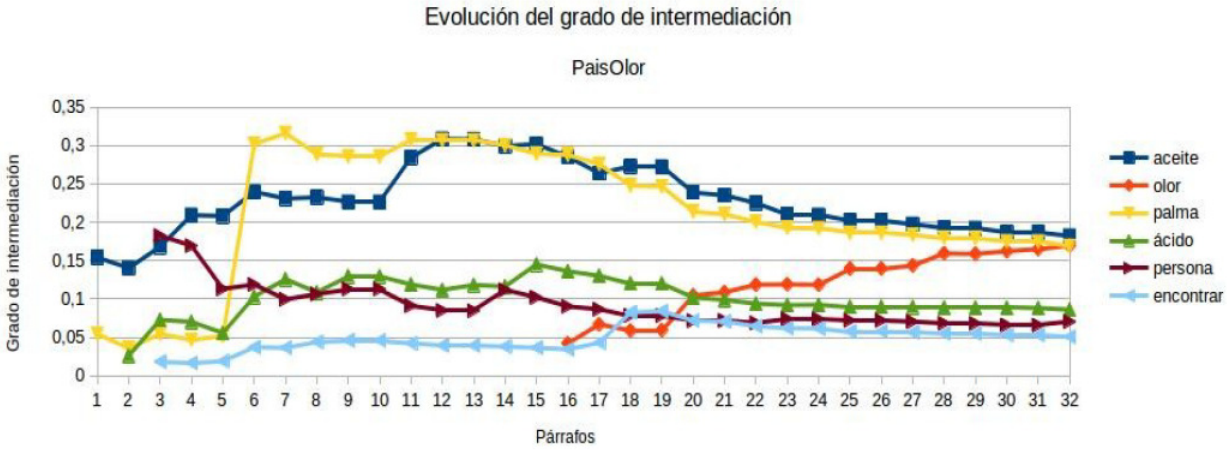

Figura 1: Evolución del grado de intermediación de las seis palabras con el grado final más alto en el texto PaisOlor

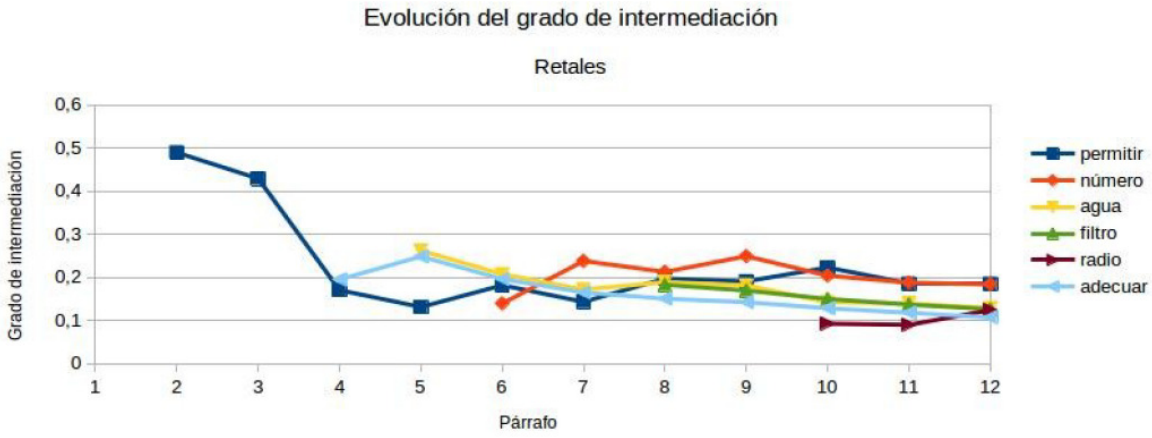

Figura 2: Evolucion del grado de intermediacion de las seis palabras con el grado final más alto en el texto *Retales*

En la Figura 1 las dos palabras con mayor intermediación –aceite y palma– van aumentando en los primeros párrafos su grado hasta que en el párrafo 16 comienzan una caída que se acentúa cada vez más; esta caída coincide con la aparición de la palabra olor que, al contrario que las otras, comienza un ascenso hasta llegar a tener casi el mismo grado que las dos anteriores.

En la Figura 2 empieza con fuerza el verbo *permitir* pero pronto empieza a caer y no hay ninguna palabra que tenga una evolución ascendente, todas presentan, un pequeño ascenso, al principio, y después, según avanza el texto, caen. Además, ninguna palabra presente en el primer párrafo acaba como una de las seis palabras con mayor grado de intermediación; las dos palabras con intermediación más alta del primer párrafo

son *plegamiento* (0,5024) y *continente* (0,3670) que acaban el texto con 0,0164 y 0,0042 respectivamente, muy por debajo de las otras y que muestran, también, un primer ascenso y, según avanza el texto, una continuada caída.

Si comparamos estos dos gráficos de evolución con los de otros dos textos –*ABC* (Figura 3) y *País* (Figura 4)– podemos diferenciar que, en los grafos de estos dos textos, frente a lo que ocurre en *PaísOlor* (Figura 1), las palabras que acaban con la intermediación más alta mantienen siempre una evolución ascendente o mantenida a lo largo del texto y la entrada de nuevas palabras no provoca una caída espectacular de su grado de intermediación; y, frente a lo que ocurre en *Retales* (Figura 2), la tendencia de las palabras es ascendente y muestran el claro predominio de algunas palabras sobre otras.

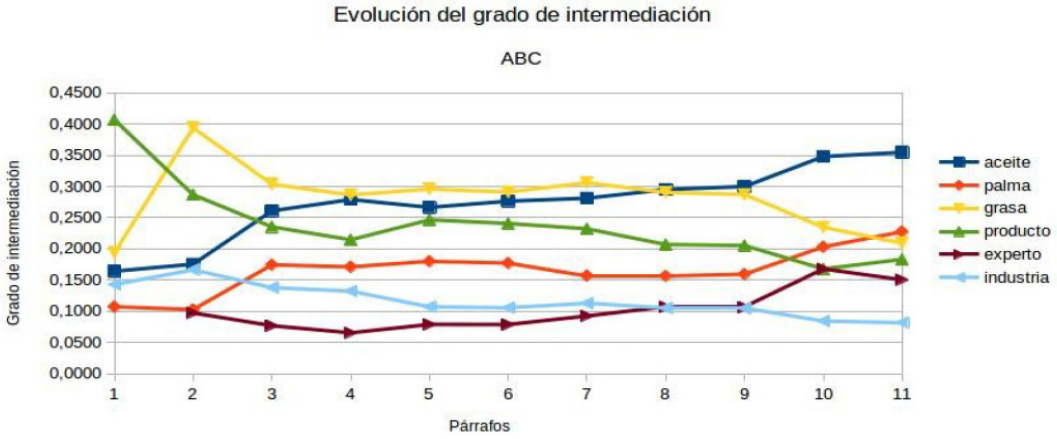

Figura 3: Evolución del grado de intermediación de las seis palabras con el grado final más alto en el texto *ABC*

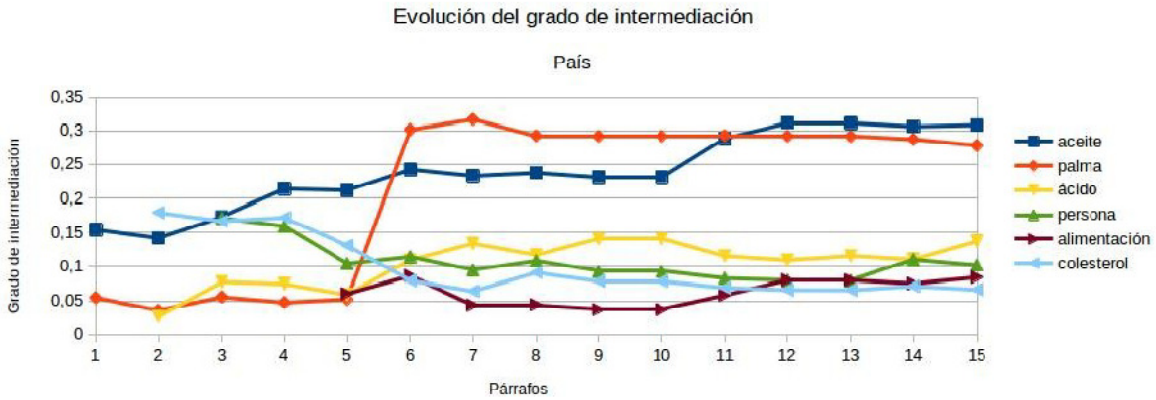

Figura 3: Evolución del grado de intermediación de las seis palabras con el grado final más alto en el texto *País*

Creo que con este análisis queda mostrado como el grado de intermediación puede ayudar a comprender la estructura temática del texto, ya que no solo permite conocer las palabras con mayor grado de intermediación, que podrían ser tomadas también como palabras clave, sino ver su evolución y detectar, a través de ellas, partes de texto poco o menos conectadas, tal como ocurre en *PaisOlor*, o textos con partes muy aisladas y poco cohesionadas. Es indudable que se podría hacer un análisis más detallado y que, en el futuro, será necesario realizar análisis de muchos más textos para poder ver las regularidades en la evolución de textos de diferente tipo y grado de cohesión; pero ahora pasemos a analizar la progresión temática de una manera más profunda a través de la evolución de las comunidades del grafo.

3.3. Progresión temática por comunidades y párrafos

Nos centraremos ahora en la modularidad de los grafos; el coeficiente de modularidad indica el grado en el que los vértices, las palabras, tienden a agruparse en clústeres o comunidades, es decir, el grado en el que se pueden diferenciar grupos de palabras con relaciones más densas que la media del grafo; estas comunidades son "groups of vertices which probably share common properties and/or play similar roles within the graph." (Fortunato, 2010, p. 3)

En nuestros seis textos el coeficiente de modularidad, siendo el máximo 1 y el mínimo 0, era relativamente alto (vid. Tabla 2) lo que indica que hay grupos de palabras que mantienen un números relaciones más alto que la media del grafo y que comparten propiedades o tienen una función similar. La

hipótesis de partida es que esa función similar que tienen las palabras de una comunidad está relacionada con aspectos semántico-temáticos, de manera que, a través de las distintas comunidades podemos averiguar las diferencias temáticas del texto.

Veamos ahora que grado de representación tiene cada comunidad en los distintos párrafos del texto. Para ello hemos hecho un cálculo muy simple, hemos contado el número de palabras del párrafo pertenecientes a cada comunidad y hemos dividido está cantidad por el número total de palabras del párrafo; esta medida nos indica la aportación de cada comunidad a cada párrafo. Hemos trasladado a un gráfico este porcentaje de aportación de palabras de cada comunidad y hemos obtenido un gráfico en el que se puede observar la progresión temática de cada texto.

Analizaremos, en primer lugar, la progresión de un texto –*ABC* (Figura 5)– y la comparemos con la de un texto de control –*Retales* (Figura 6)– para observar sus diferencias.

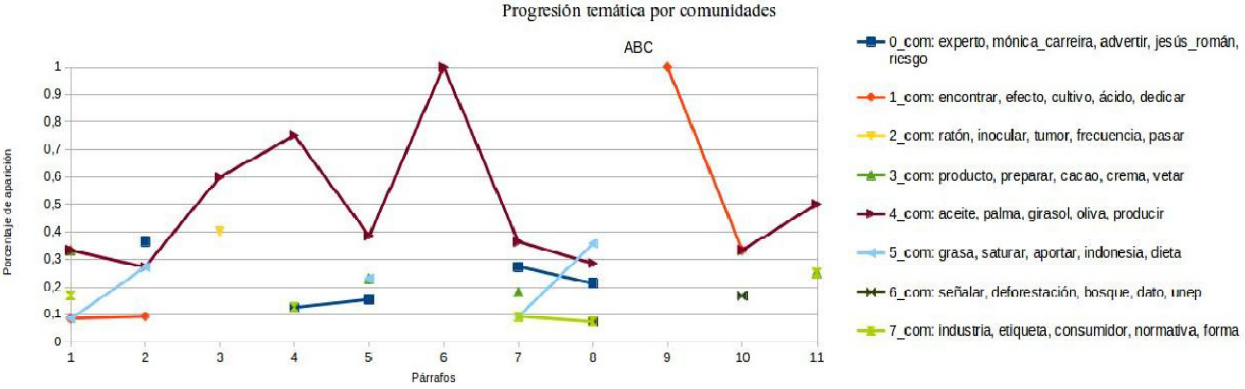

Figura 5: Progresión temática por comunidades y párrafos de *ABC*

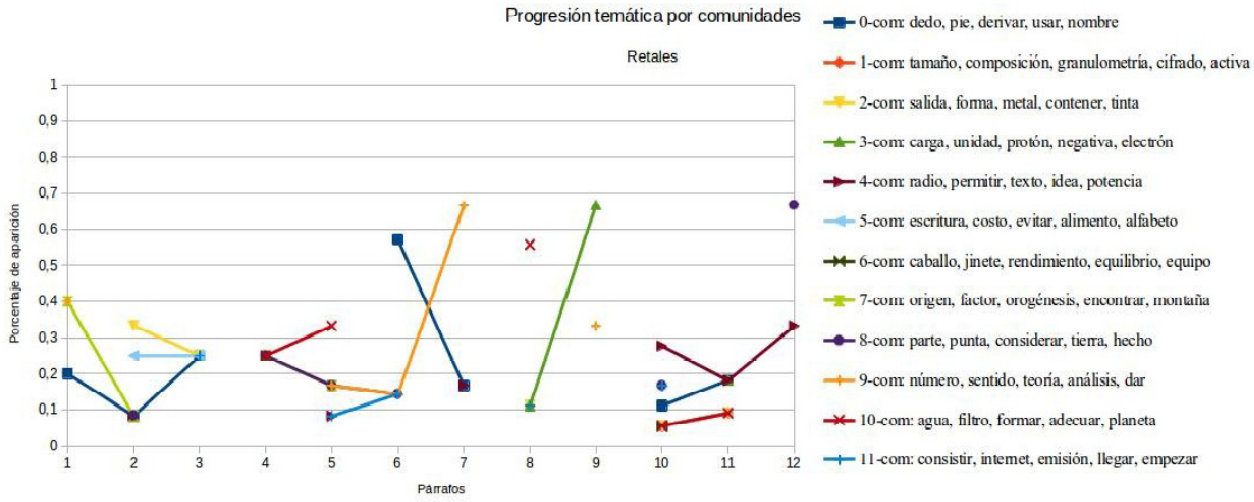

Figura 6: Progresión temática por comunidades y párrafos de *Retales*

Antes de analizar el gráfico 5, hemos de aclarar que las palabras que aparecen en la leyenda son solo las cinco palabras pertenecientes a cada comunidad con el grado (*degree*) más alto (Newman, 2010, pp. 168-169; Opsahl et al., 2010) y que han sido escogidas como sus representantes para ayudar a la comprensión del gráfico, pero hay muchas más palabras en la comunidad. En la leyenda del gráfico, se ha respetado el número de comunidad asignado por el algoritmo (Nº_com), por lo que el número no representa ningún tipo de secuencialidad textual.

Aclarado estos dos puntos, vemos que, en el texto *ABC*, aparece un tema más o menos constante (4_com) que, con variaciones, se mantiene desde el párrafo 1 al 8 y que reaparece como cierre en los dos últimos. Hay dos párrafos que contienen un solo tema –el 6, tema 4_com y el 9, tema 1_com– con valor de 1; son títulos de subapartados, párrafos breves con pocas palabras normalmente pertenecientes a una única comunidad. A lo largo del texto, se puede ver como, en mayor o menor medida, son tratados diferentes temas; por ejemplo, la opinión de los expertos, fundamentalmente asignada a 0_com, aparece en los párrafos 2, 4 y 5, y vuelve a aparecer en el 7 y en el 8; este tema está relacionado con el de la industria y la normativa (7_com) ya que aparecen conjuntamente en el 4 y en el 7 y 8. Hay, también, casos de temas que aparecen solamente una vez, el 2_com, que trata sobre experimentos en ratones, en el párrafo 3. Fijémonos, además, que, en el párrafo 1 y

2, aparecen cuatro de las siete comunidades temáticas diferenciadas en lo que sería la introducción y presentación de temas.

Comparemos ahora este gráfico con el del texto *Retales* (Figura 6, en la página anterior) que, hay que recordar, estaba formado por párrafos de temáticas diferentes y escogidos al azar. A diferencia de lo que ocurría en el anterior, no hay una comunidad constante que cohesione los párrafos, los temas aparecen y desaparecen, en algunos casos se repiten, hecho que, al ser párrafos sueltos, parecería un poco extraño; esto se debe a que muchas palabras pueden repetirse aunque el tema cambie radicalmente, por ejemplo, *agua* de 10_com o *consistir* de 11_com pueden aparecer en textos de temáticas diferentes.

La pertenencia a una comunidad es una cuestión de grado, unas palabras conforman el núcleo de una comunidad y otras la periferia, esas palabras periféricas pueden aparecer cuando se trata otro tema; una de las labores que queda por hacer es encontrar la forma de medir ese grado de pertenencia para poder ser más precisos en la descripción de la progresión temática.

Comparemos ahora la progresión en el texto *País* (Figura 7) y en el texto *PaisOlor* (Figura 8) que, recordemos, es la unión del primero con un texto sobre otro el tema del mal olor.

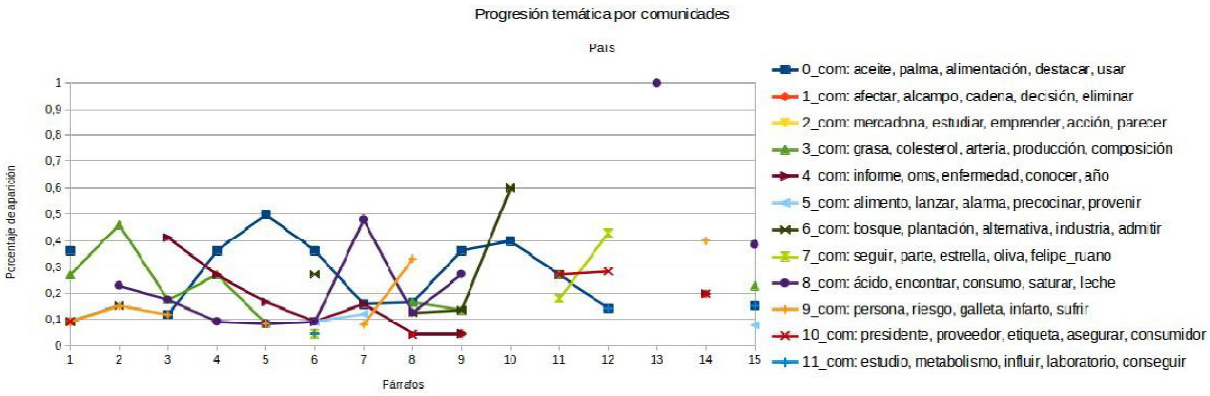

Figura 7: Progresión temática por comunidades y párrafos de *País*

En este texto, se vuelve a ver, aunque de forma menos marcada, como algunos temas discurren a lo largo de casi todo el texto (0_com, 8_com) que tratan sobre el aceite de palma, el ácido palmítico, la alimentación y el consumo. Vemos que, en los tres primeros párrafos, están presentes dos temas muy relacionados (3_com y 9_com) sobre las consecuencias para la salud (*grasa*, *colesterol*, *arteria*, *persona*, *riesgo*, *infarto*). En el párrafo 13 aparece el título de un apartado, en este caso, se trata más bien de un párrafo muy breve ("no todo el ácido palmítico es el demonio.") para, en los dos finales, retomar casi todos los temas tratados: 1_com, 4_com, 9_com y 10_com en el párrafo 14 y 0_com, 3_com, 5_com, 8_com, 11_com en el 15.

En *PaisOlor* (Figura 8), la unión de los dos textos con temáticas diferentes produce un corte en los temas en el párrafo 15, ya que, a partir de ahí no se repite casi ninguno de los temas y, los que se repiten, lo hacen de forma muy escasa.

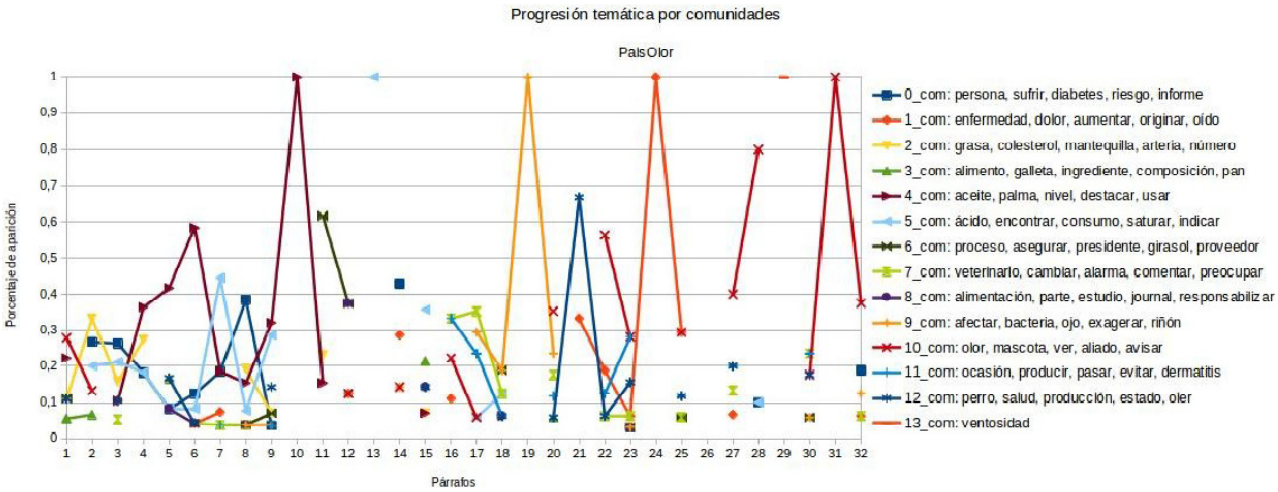

Figura 8: Progresión temática por comunidades y párrafos de *PaisOlor*

Hay que indicar que, en el gráfico, no se han unido los grafos de los dos textos, sino que se han unido los textos y después se ha realizado el análisis como si de un texto único se tratase; por eso, algunos temas se pueden repetir, ya que los textos tienen palabras en común y esas palabras unen ambos grafos y pertenecerán a comunidades presentes en ambos textos. Por ejemplo, *eliminar*, aparece en los párrafos 1 y 25, *informar*, en el 1 y 27, o *favorecer*, en el 2, 20 y 30, estas tres palabras pertenecen al 10_com, por eso este tema aparece en los dos textos. Aquí lo importante es que el corte temático aparece de forma clara entre los párrafos 15 y 16.

Por último mostrar un texto con un desarrollo diferente a *ABC* o *País*, el texto *Comidista* que parece desarrollarse por partes, no completamente aisladas pero, hasta cierto punto, independientes (Figura 9)

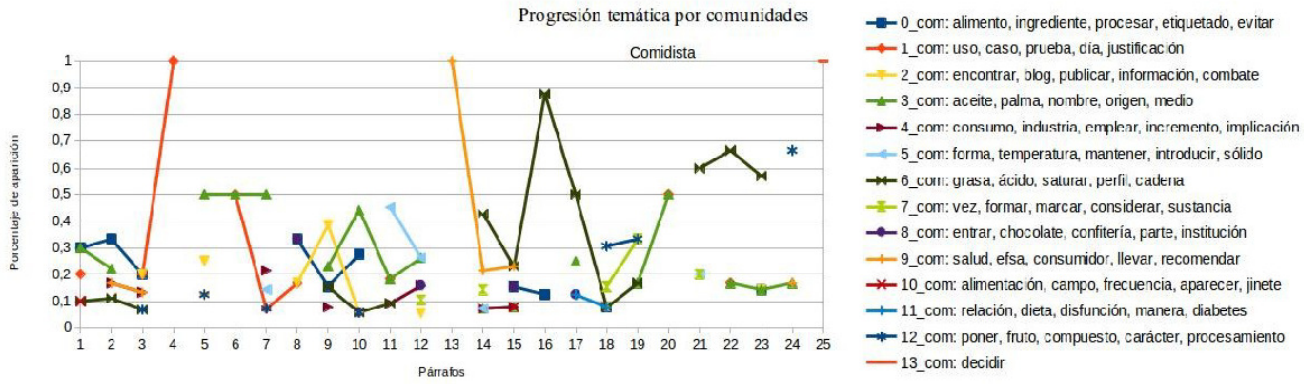

Figura 9: Progresión temática por comunidades y párrafos de *Comidista*

El párrafo 4, el 13 (ambos son títulos de apartados) y el 20 parecen marcar puntos de inflexión. Fijémonos que, en esas partes, predominan algunos temas; por ejemplo, entre el párrafo 5 y el 12, aparece el 3_com, sobre el origen del aceite, o el 2_com sobre la información y el combate contra el aceite, o el 1_com, sobre el uso. En lo que podríamos llamar tercera parte (párrafos 14-20) el tema más constante es el 6_com que habla de las cadenas de grasas saturadas; la última parte es un tanto extraña porque es una enumeración. De todo ello podemos concluir que es un texto bastante parcelado aunque en sus parcelas los temas se repiten y, también podemos extraer trabajo para el futuro, para, con muchos más textos, comprobar qué constantes numéricas tiene este tipo de textos y cuáles las enumeraciones, si es que esas constantes existen.

3.4. Análisis visual del grafo

Cuando se hace un grafo de un texto, lo normal es, primero, analizar el grafo completo, pero he preferido hacerlo al revés pues creo que, de esta manera, será más fácil comprender qué se puede extraer de la forma visual de los grafos. Analizamos ahora algunas características de los grafos completos de cuatro de los textos –*ABC*, *Comidista*, *PaísOlor* y *Retales*– para comprobar qué podemos intuir de su forma. Hay que indicar, antes, que todos los grafos (Figuras 10 a 12, en las páginas siguientes) tienen una distribución *ForceAtlas* con una repulsión de 20000, el tamaño de las etiquetas de los nodos es proporcional a su grado de indeterminación y el color de los nodos corresponde a las distintas comunidades.

En la Figura 10 y en la 11, se puede ver como dos textos con distinta estructura presentan un grafo diferente. En la primera, el texto *ABC*, presentaba dos temas claros que eran constantes a lo largo del texto, constituyen, en el grafo, un núcleo bien definido del que parten los otros temas como brazos de una estrella; en la segunda figura, el texto *Comidista*, se había visto que no tenía temas tan constantes, de ahí que no tenga un núcleo tan claro y que los posibles brazos, a diferencia del anterior, aparezcan menos nítidos y más embebidos en el disperso núcleo central.

Por otro lado, en la Figura 12, el grafo del texto *PaísOlor,* creado por la unión de dos textos, muestra dos claros núcleos con las palabras *olor* y *aceite* como nodos centrales de cada uno; y, para finalizar, un grafo sin núcleo claro o con muchos pequeños núcleos es el que aparece en la Figura 13 que corresponde al texto *Retales*, con sus brazos no claramente relacionados con ningún núcleo, lo esperable de un texto creado por párrafos escogidos al azar de temáticas muy dispares entre sí.

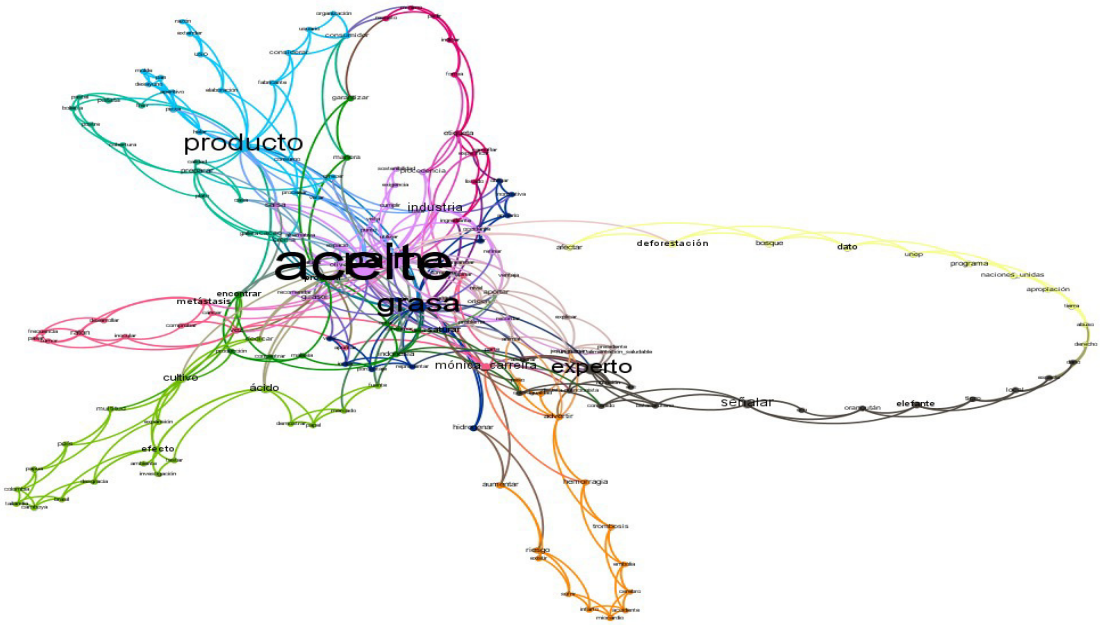

Figura 10: Grafo de *ABC*

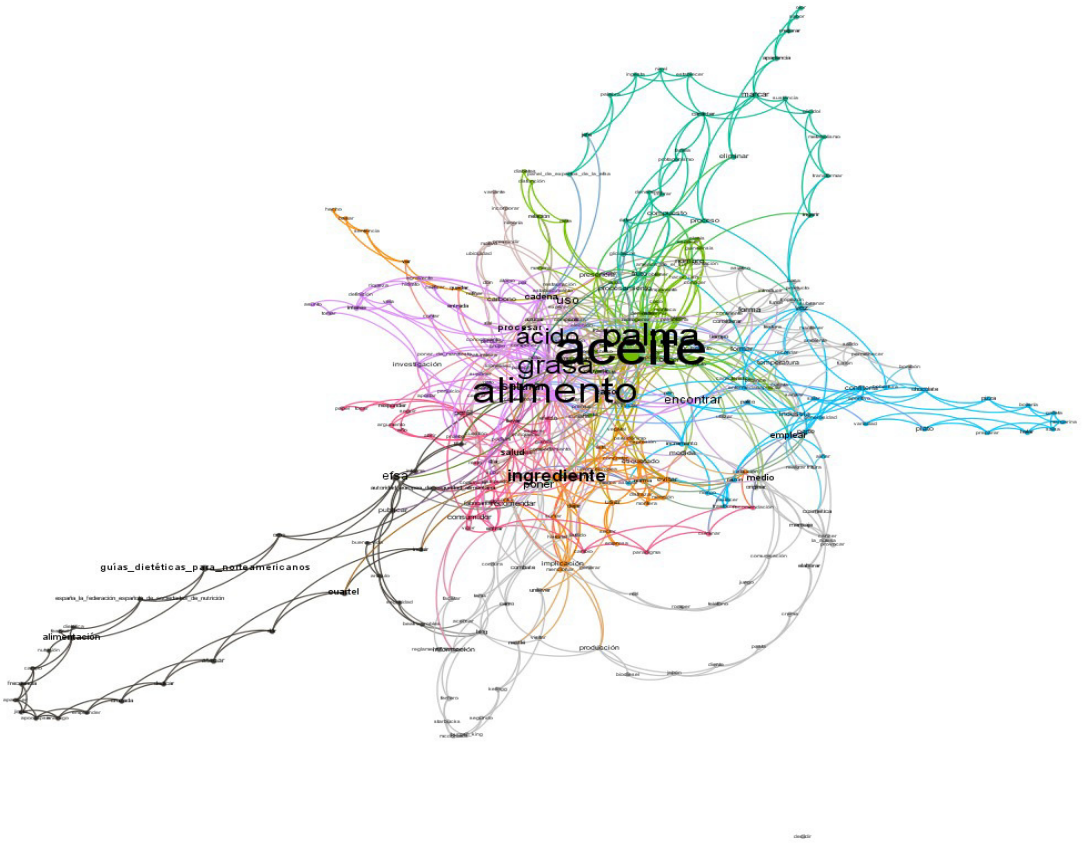

Figura 11: Grafo de *Comidista*

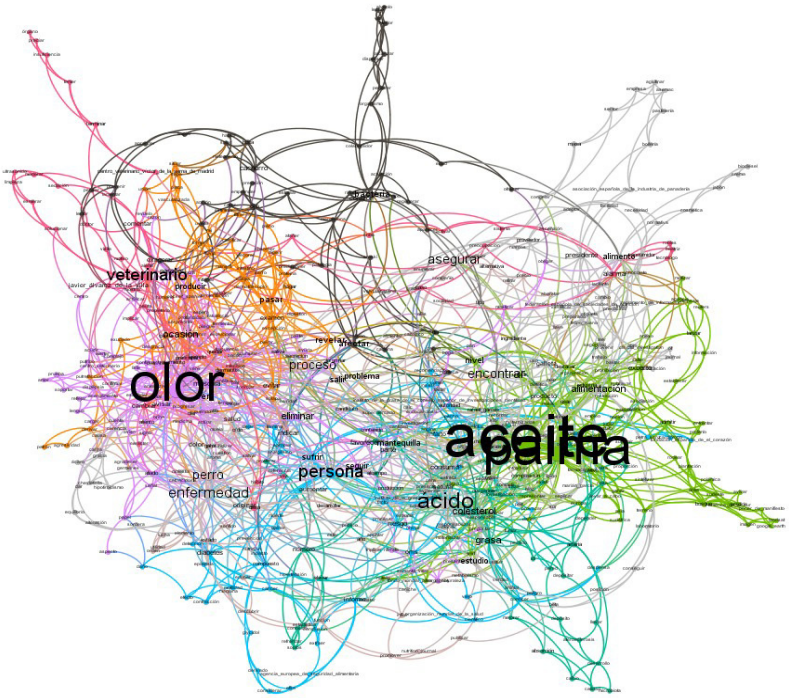

Figura 12: Grafo de *PaisOlor*

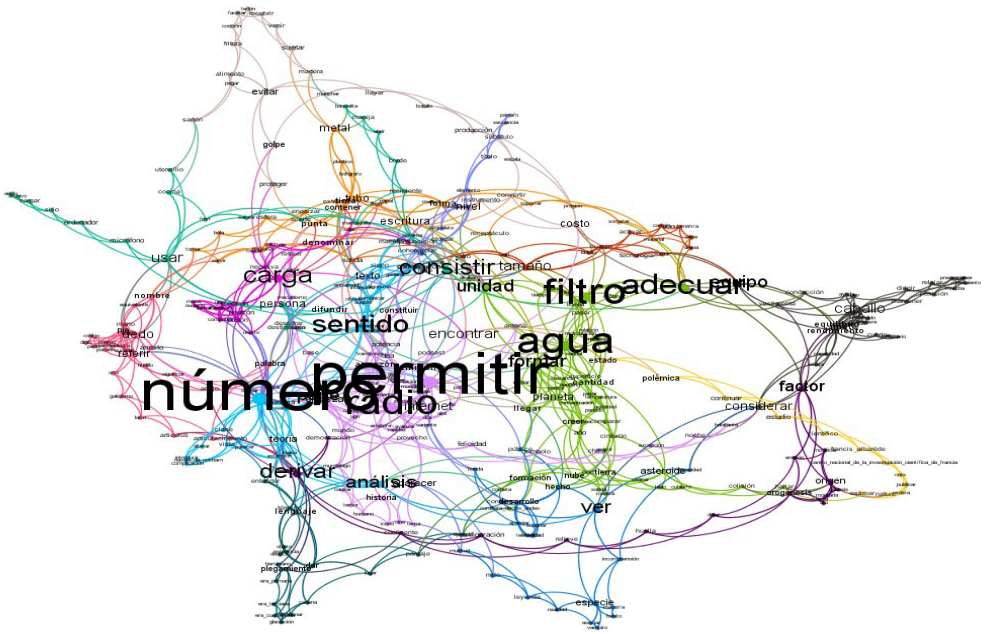

Figura 13: Grafo de *Retales*

4. Conclusiones

Se han mostrado las bondades de la aplicación de la teoría de grafos y de sus métodos en el estudio de la progresión temática y se han apuntado, también, algunas potencialidades de su aplicación en el análisis de la cohesión léxica, los objetivos generales propuestos al comienzo del trabajo.

El uso de la intermediación establece los temas principales del texto como unidad global; el análisis de la evolución de esta medida a lo largo del texto revela como se distribuyen los temas principales en los párrafos.

El uso de la modularidad y el estudio de las comunidades generadas por esta permiten analizar la contribución de las palabras de cada comunidad a cada párrafo y poder determinar en qué párrafos son tratados los distintos subtemas textuales mostrando su progreso en el proceso de escritura, en el texto visto como progresión temporal.

La aplicación de ambas medidas facilita la discriminación de textos más y menos cohesionados temáticamente y ayuda a la identificación de su estructura temática.

Es cierto que aun quedan muchos aspectos por estudiar, algunos de ellos ya apuntados: comprobar la potencialidad del método en otros tipos de textos (poéticos, narrativos, descriptivos…); estudiar si existen diferencias topológicas en textos de distinto tipo; medir las diferencias entre palabras periféricas y centrales en las comunidades y las aplicaciones que estas diferencias puedan tener; comprobar si, a través de los grafos, se puede llegar a diferenciar distintos despliegues expositivos (enumeraciones, secuencias temporales, etc.); determinar si la forma global del grafo se corresponde con distintos despliegues temáticos y si es posible establecer una tipología de formas que ayude a identificar visualmente y de forma rápida los tipos de despliegue.

Todos estos aspectos, todavía no estudiados, y muchos más creemos que no disminuyen el valor del estudio, sino que, al revés, abren un campo de investigación amplio y prometedor.

Corpus utilizado

¿Es realmente malo el aceite de palma? (s. f.). Recuperado 9 de mayo de 2017, de http://www.abc.es/sociedad/abci-realmente-malo-aceite-palma-201702211242_noticia.html

¿Malos olores? (2017, abril 5). abc. https://www.abc.es/sociedad/abci-malos-olores-201704052220_noticia.html

Pérez, C. (2015, febrero 2). *Beneficios de comer aceite de oliva crudo*. Natursan. https://www.natursan.net/beneficios-de-comer-aceite-de-oliva-crudo/

¿Por qué Alcampo quiere retirar el aceite de palma de sus productos, si no es tóxico ni venenoso? | BuenaVida | EL PAÍS. (s. f.). Recuperado 9/05/2017, de http://elpais.com/elpais/2017/04/04/buenavida/1491318026_847822.html

¿Por qué es malo el aceite de palma? | El Comidista EL PAÍS. (s. f.). Recuperado 9 de mayo de 2017, de http://elcomidista.elpais.com/elcomidista/2017/02/16/articulo/1487259154_419212.html

Referencias bibliográficas

Alcíbar Cuello, J. M. (2004). *La divulgación mediática de la ciencia y la tecnología como recontextualización discursiva*. https://idus.us.es/xmlui/handle/11441/24760

Amancio, D. R. (2015). A Complex Network Approach to Stylometry. *PLOS ONE*, *10*(8), DOI: 10.1371/journal.pone.0136076

Antiqueira, L., Nunes, M. das G. V., Oliveira Jr, O., & F Costa, L. da. (2007). Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, *373*, 811–820. https://arxiv.org/pdf/physics/0504033.pdf

Antiqueira, L., Nunes, M., Oliveira Jr, O., & Costa, L. da F. (2005). Modelando textos como redes complexas. *Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana*, 22–26. http://nilc.icmc.usp.br/til/til2005/arq0054.pdf

Barranco Flores, N. (2015). Las cadenas nominales y la estigmatización de la realidad referida en el periodismo informativo. En S. Henter, S. Izquierdo, & R. Muñoz (Eds.), *Estudios de pragmática y traducción* (pp. 119-134). EDITUM (Ediciones de la Universidad de Murcia).

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, 111–121. DOI: 10.7916/D85B09VZ

Bastian, M., Heymann, S., Jacomy, M., & others. (2009). Gephi: An open source software for exploring and manipulating networks. *ICWSM*, *8*, 361–362. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154/1009/

Bernárdez, E. (1982). *Introducción a la lingüística del texto* (Vol. 1). Espasa-Calpe Madrid.

Boudin, F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. *arXiv preprint arXiv:1803.08721*. https://arxiv.org/pdf/1803.08721.pdf

Boudin, F. (2013). A comparison of centrality measures for graph-based keyphrase extraction. *International Joint Conference on Natural Language Processing (IJCNLP)*, 834–838.

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. *International Joint Conference on Natural Language Processing (IJCNLP)*, 543–551. https://hal.archives-ouvertes.fr/hal-00917969/document

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, *25*(2), 163-177. https://doi.org/10.1080/0022250X.2001.9990249

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge University Press.

Calsamiglia Blancafort, H., & Tusón Valls, A. (2004). *Las cosas del decir. Manual de análisis del discurso*. Ariel.

Calsamiglia, H., & Van Dijk, T. A. (2004). Popularization discourse and knowledge about the genome. *Discourse & society*, *15*(4), 369–389. DOI: 10.1177/0957926504043705.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).*, 239–242. https://www.researchgate.net/profile/Muntsa_Padro/publication/228976391_Freeling_An_open-source_suite_of_language_analyzers/links/02bfe50fd836ec3df3000000.pdf

Cassany, D. (1995). *La cocina de la escritura*. Anagrama.

Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Ding, J. (2018). *Linguistic prefabrication*. Springer Berlin Heidelberg.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management*, *43*(6), 1705-1714. DOI: 10.1016/j.ipm.2007.01.015

Ferrer I Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proc. Biol. Sci.*, *268*(1482), 2261-2265. DOI: 10.1098/rspb.2001.1800

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

García Berrio, A., & Albaladejo Mayordomo, T. (1983). Estructura composicional: Macroestructuras. *ELUA. Estudios de Lingüística, N. 1 (1983); pp. 127-179*.

Garg, M., & Kumar, M. (2018). Identifying influential segments from word co-occurrence networks using AHP. *Cognitive Systems Research*, *47*, 28-41. https://doi.org/10.1016/j.cogsys.2017.07.003

Grineva, M., Grinev, M., & Lizorkin, D. (2009). Extracting key terms from noisy and multitheme documents. *Proceedings of the 18th international conference on World wide web*, 661–670.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hirst, G., St-Onge, D., & others. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, *305*, 305–332. https://books.google.es/books?hl=es&lr=&id=Rehu8OOzMIMC&oi=fnd&pg=PA305&dq=Lexical+chains+as+representations+of+context+for+the+detection+and+cor-+rection+of+malapropisms&ots=Iqp9NmUPg7&sig=Oz2dPzaF2ucbMCsIw7kS8KGRBss

Lahiri, S. (2013). Complexity of word collocation networks: A preliminary structural analysis. *arXiv preprint arXiv:1310.5111*.

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*. https://arxiv.org/pdf/0812.1770

Landauer, T. K. (Ed.). (2007). *Handbook of latent semantic analysis*. Erlbaum.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Mapelli, G. (2006). Estrategias lingüístico-discursivas de la divulgación científica. *Scrittura e conflitto: Actas del XXI Congreso Aispi= Atti del XXII Convegno Aispi: Catania-Ragusa 16-18 mayo*, 169–184. http://dialnet.unirioja.es/servlet/articulo?codigo=2352360&orden=131167&info=link

Martínez Caro, E. M. (2014). El párrafo como unidad discursiva: Consideraciones de forma y contenido relativas a su demarcación y estructuración. *Estudios de lingüística del español*, *35*, 189-213. http://infoling.org/elies/35/elies35.1-8.pdf

Mihalcea, R., & Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

Morris, J., Beghtol, C., & Hirst, G. (2003). Term relationships and their contribution to text semantics and information literacy through lexical cohesion. *Proceedings of the 31st Annual Conference of the Canadian Association for Information Science*, 153-168. CiteSeerX.psu:10.1.1.84.3611

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, *17*(1), 21–48. http://dl.acm.org/citation.cfm?id=971740

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245-251. DOI: 10.1016/j.socnet.2010.03.006

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. *LREC2012*. http://hdl.handle.net/2117/15986

Paranyushkin, D. (2011). Identifying the pathways for meaning circulation using text network analysis. *Nodus Labs, Berlin*. http://noduslabs.com/publications/Pathways-Meaning-Text-Network-Analysis.pdf

Siepmann, D., & Siepmann-Gallagher-Hannay-Mackenzie (Eds.). (2008). *Writing in English: A guide for advanced learners*. Francke.

Solé, R. V., Corominas-Murtra, B., Valverde, S., & Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity*, *15*(6), 20–26. http://onlinelibrary.wiley.com/doi/10.1002/cplx.20305/abstract

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. En T. K. Landauer (Ed.), *Handbook of latent semantic analysis* (Vol. 427, pp. 424–440). Erlbaum.

Stubbs, M. (2001). Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence. En *The handbook of discourse analysis* (pp. 304-320). Blackwell Publishers.

Tanskanen, S.-K. (2006). *Collaborating towards coherence: Lexical cohesion in English discourse*. John Benjamins Pub. Co.

Venegas, V. (2003). Análisis semántico latente: Una panorámica de su desarrollo. *Revista signos*, *36*(53), 121–138. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342003005300008

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440–442. DOI: 10.1038/30918

Zhou, Z., Zou, X., Lv, X., & Hu, J. (2013). Research on weighted complex network based keywords extraction. *Workshop on Chinese Lexical Semantics*, 442–452. DOI: 10.1007/978-3-642-45185-0_47