APLICACIÓN DE HERRAMIENTAS DE LINGÜÍSTICA COMPUTACIONAL EN FOROS VIRTUALES

Paz Ferrero
paz.ferrero@uam.es
Facultad de Filosofía y Letras - Universidad Autónoma de Madrid

Javier Alda
j.alda@opt.ucm.es
Escuela Universitaria de Óptica - UCM

Agradecimientos: Los autores de este trabajo desean expresar un agradecimiento especial a Ana M.ª Fernández-Pampillón y a sus alumnos de la enseñanza «Lingüística Computacional». A la vez, este trabajo habría sido imposible sin la dedicada participación de los Coordinadores de Centro de CV-UCM, quienes con sus intervenciones han ofrecido un material de gran calidad y variedad.

Palabras clave[1]: lingüística computacional, foros virtuales, palabras clave, anafóricos.

Se ha aplicado una herramienta automática de análisis de textos, *Wordsmith*, para la obtención de información cuantitativa acerca del contenido de foros en un entorno de enseñanza virtual sin requerir la intervención humana en su lectura. Se ha podido establecer el carácter formal del discurso y se han identificado participantes que utilizan contenidos léxicos diferenciados. Se han creado lemas específicos para la extracción automática de elementos gramaticales y semánticos. La identificación automática de palabras clave ha permitido validar cuantitativamente los resultados obtenidos.

INTRODUCCIÓN

Dentro del desarrollo de la lingüística computacional, en los últimos años han proliferado herramientas automáticas capaces de analizar textos de forma eficaz y cuantitativa en todos los niveles del lenguaje. Incluso, dependiendo de la herramienta informática, se puede inferir y reconstruir gran parte del significado de todo un texto. No obstante, el manejo y aprovechamiento de las prestaciones de una herramienta automática requieren un adiestramiento previo y un conocimiento mínimo de convencionalismos informáticos y lingüísticos y de conceptos semánticos (Landauer, 1999). El procesamiento automático de textos nos permite una gran rapidez en la obtención de datos cuantificables, pero la validez y fiabilidad de esos datos dependerán de factores tales como el rendimiento operativo de la herramienta, la disponibilidad de un *corpus* textual adaptado, el nivel de preprocesado de los textos, la calidad y configuración de un *lemario* o *lexicón específico,* el análisis manual de los datos obtenidos y, por último, la comprobación e interpretación de los resultados.

Dentro de las numerosas herramientas automáticas de análisis cuantitativo, bien comerciales o de libre distribución, podemos mencionar *TACT, Análisis 2.94, AntConc 3.1.302, AntText,* entre otras. Nosotros hemos optado

[1] Las palabras clave seleccionadas (excepto «lingüística») han sido extraídas a partir de las computadas con *Wordsmith,* tomando como referencia léxica el *corpus* de los foros estudiados en este trabajo.

por la herramienta comercial *Wordsmith,* por su sencillez en el manejo y sus prestaciones. Además, se presenta como una herramienta tan apta para la enseñanza como para la investigación. Según el experto lingüista computacional Mark Davies, autor de interesantes *corpora* de textos literarios en castellano, considera que *Wordsmith* «es una herramienta con la que se puede hacer muchas millas» (Davies, 2006).

MOTIVACIÓN

Nuestro interés por este trabajo es múltiple. Por un lado, *Wordsmith* nos permite clasificar, cuantificar y contextualizar el vocabulario al mismo tiempo que registra la frecuencia y observar las peculiaridades de las palabras utilizadas. Además, nos interesa comprobar que, a partir del tratamiento automático de textos, no sólo podemos cuantificar y analizar el léxico, sino obtener también una idea aproximada del contenido semántico de dicho texto. El objetivo último sería el análisis automático de textos generados en foros virtuales, la caracterización de los mensajes y la extracción de los temas generales tratados en dichos foros mediante la localización de las palabras más frecuentes, palabras clave y sus combinaciones, tanto en el «asunto» como en el «cuerpo» del mensaje. De ahí que con un estudio de este tipo podamos reconstruir parcialmente el significado de los mensajes y caracterizar tanto la estructura de los mensajes como su estilo expresivo sin requerir la lectura completa de los textos.

Además, mostraremos cómo una secuencia de mensajes pertenecientes al mismo «asunto» puede mantener la coherencia a lo largo de toda la cadena de intervenciones o, por el contrario, desvirtuar el contenido del discurso independientemente de la longitud de la cadena de mensajes. Esto es, se puede mantener el mismo «asunto», pero se puede ir desvirtuando su tema bien desde el estado inicial o bien a lo largo de la cadena. Este tipo de discurso inconexo ilustra un fenómeno conocido como *pérdida de coherencia* (Landauer, 1999; Dong, 2004). En nuestro caso, la falta de coherencia la acuñaremos con el término de *discurso 1/f* y lo ejemplificaremos más abajo. Con esta observación no se trata de valorar negativamente esta divergencia o falta de correlación progresiva de los mensajes, sino de apuntar que esta característica, propia de la conversación, también puede darse en un foro de debate.

MATERIAL

Los textos que analizamos son los mensajes generados en los foros principales de debate en el espacio denominado «Coordinadores uCm» a lo largo de tres cursos: 2003-2004, 2004-2005 y 2005-2006 por los coordinadores de Campus Virtual de la Universidad Complutense de Madrid (CV-UCM) y dinamizados por los miembros de la UATD-CV (Unidad de Apoyo Técnico y Docente al Campus Virtual UCM). Este grupo de profesores da soporte a sus respectivos centros en todos los temas relacionados con las nuevas tecnologías de las comunicaciones aplicadas a la enseñanza. Tanto la naturaleza de los textos como su disposición en cadena tienen unas peculiaridades determinadas por la temática y el soporte en sí. Es decir, trabajamos con unos textos de comunicación virtual producidos en unos foros de discusión de un entorno educativo de la plataforma WebCT.

Además, los mensajes objeto de nuestro estudio se caracterizan por tener una estructura genérica: *Asunto* y *cuerpo*. A su vez, el *cuerpo* generalmente se estructura en *saludo*, *información-aportación* (con o sin mensaje anterior citado) y *despedida*. Esta estructuración sistemática nos va a permitir agrupar en un texto único todos los foros y tratar los textos de forma sencilla y selectiva según las necesidades del estudio. El trabajar con mensajes generados en la plataforma *WebCT* tiene varias ventajas. Por un lado, su herramienta de «Buscar» nos permite acceder a mensajes concretos según diferentes criterios: palabra clave, autor, fechas, asunto, etc., como puede verse en la figura 1. De hecho, algunas comprobaciones y datos preliminares de este trabajo se han realizado utilizando esta herramienta (recuento de mensajes por días de la semana, por autores, etc.).

Por otro lado, esta plataforma permite descargar los mensajes cómodamente a ficheros de texto plano, formato típicamente utilizado en la mayoría de herramientas automáticas, en nuestro caso *WordSmith*.

Figura 1. Herramienta de selección de mensajes de WebCT

METODOLOGÍA

La materia prima objeto de estudio se ha agrupado en un texto único y principal al que denominamos *corpus*. Este *corpus* contiene todos los mensajes de los foros y se concibe como el texto de referencia para computar automáticamente unidades de texto menores. A partir de este *corpus* cuantificaremos tanto las *palabras contenido* (nombres, verbos, pronombres, adjetivos y adverbios) como las *palabras con función* (preposiciones y las conjunciones) (Just & Carpenter, 1987:139) mediante la herramienta cuantitativa *Wordsmith* con sus tres aplicaciones fundamentales: *Wordlist* para realizar el listado de palabras, *Keywords* para hallar la *palabra clave* y *Concord* para ubicar la palabra en su contexto.

En primer lugar, a partir del *corpus,* hemos generado un listado de palabras con sus frecuencias de aparición. A continuación, para minimizar la extensión del listado, hemos agrupado de forma automática las *palabras con contenido y las palabras con función* o *gramaticales*. El listado de *palabras con contenido* se ha obtenido porque previamente hemos indexado en una misma palabra aquellas que tienen un mismo lexema o un significado similar. De este modo, hemos utilizado dos criterios para agrupar las palabras: el morfológico y el semántico (Pressley & Afflerbach, 1995). Desde el punto de vista morfológico, agrupamos las palabras con significado similar según su diferente flexión, composición, derivación o afijación. Desde el punto de vista semántico, hemos agrupado palabras con similar significado aplicando los conceptos de hiperonimia-hiponimia (relación jerárquica entre las palabras), sinonimia (palabras distintas que *connotan* un significado semejante) e «invocación» (palabras diferentes pero que *denotan* un mismo concepto o entidad).

Esta doble concepción en la reagrupación por significante y significado nos permite obtener un listado muy completo y específico de cuantificación, e incluso agrupar categorías gramaticales (adverbios, preposiciones, elementos anafóricos y pronombres…). Por ello, para registrar el máximo número de palabras, hemos creado unas palabras genéricas (anaforicospfg, discursoemotivopfg, adverbiospfg, etc.) para aglutinar las palabras asociadas y poder cuantificarlas automáticamente mediante *Wordsmith*.

Una vez obtenida una sola lista o lemario con los datos clasificados seguimos un proceso deductivo. Esta lista específica de palabras, creada *ex profeso* para este estudio, empezará a darnos información básica de los foros. Por ejemplo, podremos observar quiénes son los participantes más activos o populares en los foros, qué tiempos verbales concurren, qué tono tiene el foro, qué días se hacen más aportaciones, etc. Incluso el lemario nos va a permitir caracterizar el talante y tono general de los foros mediante la observación y organización de la terminología usada. Toda esta identificación del perfil del foro lo realizamos de forma automática, sin realizar ninguna lectura, mediante la aplicación de *Wordlist*.

En una segunda fase nos centramos en las *palabras clave*. En nuestro análisis hemos utilizado las cuatro cadenas de discusión más largas de los tres foros. Nuestra hipótesis es que las palabras clave se encuentran en el «asunto» del mensaje. Para validar esta hipótesis hemos realizado dos procesamientos: uno completo del «asunto» y el «cuerpo» de cada cadena, y otro en el que se ha analizado únicamente el «cuerpo». La palabra clave se obtiene automáticamente comparando el texto que se desea analizar con un *corpus de refe-*

rencia específico y afín al estudio (Fernández-Pampillón, 79 y ss.). Esta operación se realiza mediante la aplicación de *Keywords*.

Por último, utilizamos la aplicación *Concord*. Con ella se puede rastrear la aparición de las *palabras clave* y sus combinaciones en su contexto, los elementos *anafóricos* y sus referentes y el entorno de algunas palabras poco frecuentes pero significativas. En nuestro caso, *Concord* ha sido útil para observar en su contexto a las palabras que hemos denominado «emotivas». Además, con *Concord* hemos creado un índex concreto con la sintaxis específica de aquellas palabras combinadas que nos interese estudiar. Esta aplicación es la que más información nos aportará, ya que la palabra objeto de análisis aparece en su propio contexto.

RESULTADOS Y DISCUSIÓN

La aplicación de *Wordlist:* listado de palabras

Después de analizar todo el *corpus* de mensajes con las distintas aplicaciones, se han obtenido datos que permiten inferir información básica de los foros. El siguiente párrafo expresa una primera interpretación, o perfil, del contenido de los foros, realizada por el autor que no ha participado en los mismos y sin haberlos leído con anterioridad al análisis (se han indicado las frecuencias de aparición de las palabras entre corchetes):

La mayoría de los *mensajes* [932] *y* [1393] *correos electrónicos* [28] *se* [597] *han* [93] *enviado* [759] durante los meses de *septiembre* [118] y *octubre* [177], meses de mayor actividad, para tener un merecido *relax* [1] los meses de *julio* [17] y *agosto* [1]. Otra *información* [63] que nos aporta la plataforma *virtual* [189] de *WebCT* [128] es que los *profesores* [286] parecen estar más preocupados por cuestiones de gestión *tecnológica* [213] para facilitar el *acceso* [94] a los *alumnos* [561] y el *apoyo* [241] a la *docencia* [227] que por *métodos* [22] *didácticos* [1] o de *enseñanza* [13]. Aunque no es un asunto muy tratado el *absentismo* [25], se apunta como problema que puede surgir al disponer de un *campus virtual* [369].

La reconstrucción de tal interpretación se ha basado en que la mayoría de las palabras más frecuentes se han asociado a grupos predeterminados, o «genericospfg», creados para agrupar el máximo de palabras afines (Graesser *et al.,* 1997). Para ello nos hemos basado en distintas concepciones de clasificación: hiperonimia-hiponimia, derivación, sufijación, sinonimia, categoría gramatical, etc. Un ejemplo de esta agrupación o *lematización* se observa en las figuras 2 y 3.

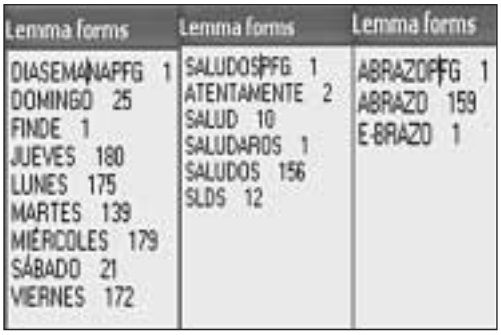


Figura 2. Lemario con los días de la semana y tipos de despedida extraídos de los foros

Figura 3. Lemario de pronombres

La frecuencia de las palabras y su reagrupación nos permite, por ejemplo, conocer los días en que se envían más mensajes al foro o las palabras más utilizadas para despedirse (fig. 2). También es llamativa la correlación pronominal. Se observa que para dirigirse los coordinadores entre sí lo hacen de forma muy interpersonal y con conciencia de grupo como muestra el uso frecuente de *yo*/*mí*/*me*-*nos*-

otros/nos-vosotros/os (fig. 3), frente al resto de las demás personas gramaticales.

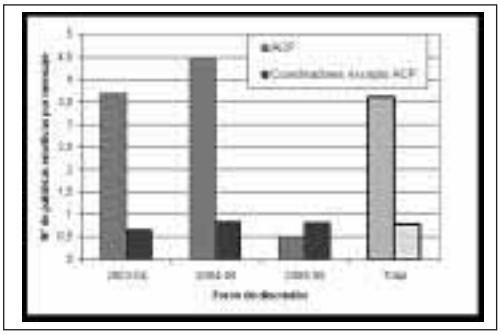


Figura 4. Comparativa entre el n.º de palabras «emotivas» de ACP y del resto de participantes

Incluso al inicio del estudio de los foros con *Wordlist* apreciamos que uno de los participantes, ACP, destacaba por su índice de participación. Es más, hemos observado que, por su especificidad léxica en el discurso, muchas de sus palabras quedaban fuera del agrupamiento general al formar parte de un léxico que calificamos como «emocional» (Ferrero & Alda, 2005). Para comprobar de forma cuantitativa esta peculiaridad se han seleccionado los mensajes de ACP, y se ha realizado un análisis detallado e individualizado de su vocabulario para después cotejarlo con el resto de los participantes. A la sazón, se ha construido un *lemario* concreto para filtrar e identificar aquellas palabras que hemos considerado con mayor carga emocional, excluyendo de este análisis los *emoticones*. Este lemario contiene diminutivos, aumentativos, palabras coloquiales, onomatopeyas, palabras relacionadas con opiniones, juicios de valor, sentimientos, sensaciones, etc. El resultado del análisis se aprecia en la figura 4.

Al analizar los datos totales de los tres foros se ha obtenido que el número promedio de palabras por mensaje de ACP (92,48 palabras) es ligeramente inferior al número promedio del resto de coordinadores (110,94 palabras). Además, el porcentaje de palabras emotivas por mensaje de ACP ha sido 5,5 veces mayor que el resto de participantes. Incluso cuando hemos estudiado los *emoticones* de los foros observamos que la abundancia de *emoticones* optimistas ☺ y guiños ;-) son más frecuentes y más utilizados por ACP que los que expresan tristeza ☹. Al ser cuantificados y comprobados en su contexto con *Concord* obtenemos, por un lado, los siguientes resultados:

Tabla I. Frecuencia de *emoticones* en los foros

Emoticones	*ACP*	*Resto de coordinadores*	*Total*
☺	53	41	94
☹	2	6	8
;-)	20	11	31

Por otro lado, cuando hemos observado el contexto de ☹ con *Concord*, éste ha encontrado dicho *emoticón* asociado a palabras o situaciones negativas como: 1- «*mala noticia* :(« 2- «*Siento el error* :(« 3- «*SÓLO FUNCIONA CON Explorer* :((:» 4- «*No hemos convocado todavía las próximas reuniones por áreas* :((« 5- «*Nos hemos incorporado todos* :((*con un poco de depresión»* y 6- «*a las 10 tengo clase* :(.» Sólo dos emoticones ☹ son de ACP. A partir de estos resultados se puede concluir que ACP tiene un discurso positivo y optimista. Además, este ejemplo muestra que la alta frecuencia de las palabras no es el único parámetro para extraer conclusiones definitivas, y que sólo un análisis más detallado de la baja frecuencia de algunas palabras puede aportar información relevante.

La aplicación de *Keywords*: identificación de la palabra clave

Esta herramienta ha resultado útil para hallar la *palabra clave* y, a su vez, conocer los asuntos de los mensajes o temática de los foros. Hemos supuesto que el «asunto» y el «cuerpo» del mensaje son coherentes temáticamente. La aplicación de *Keywords* se ha utilizado para identificar *las palabras clave* y su frecuencia tanto en el mensaje completo (asunto + cuerpo) como en el mensaje sin «asunto». En las gráficas de la figura 5 apre-

ciamos la singularidad y diferencia temática de las *palabras clave* del *«cuerpo»* frente a *las palabras clave* del mensaje completo. Para el estudio hemos elegido las cuatro cadenas más largas de cada foro y las hemos procesado junto con el *corpus*.

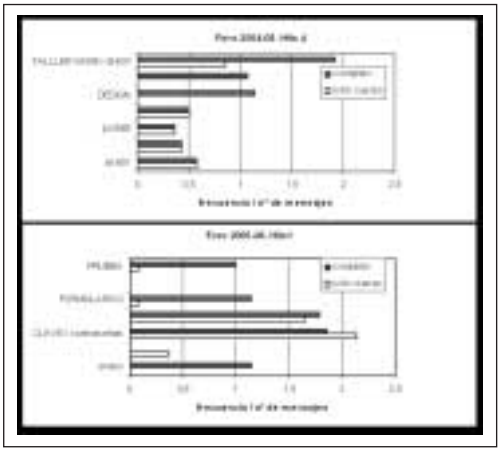


Figura 5. Frecuencias normalizadas al número de mensajes para las palabras clave *extraídas con* Keywords *al aplicarse al texto completo (en oscuro) y sólo al cuerpo del mensaje (en claro). Las palabras del asunto se muestran en mayúsculas y las palabras del cuerpo en minúsculas. Los textos analizados corresponden a dos hilos completos de 14 mensajes cada uno*

En la tabla II se muestra el resultado pormenorizado de las *palabras clave* registradas de los cuatro hilos más largos de cada uno de los foros. Al procesarlos comprobamos que, efectivamente, las *palabras clave* coinciden más con las palabras del «asunto» del mensaje que con las del «cuerpo». No obstante, este procedimiento nos permite tener una primera aproximación de la temática de los foros sin haberlos leídos previamente. Una interpretación más precisa de la temática de los mensajes se obtiene al identificar las *palabras clave* en el contexto del «cuerpo» y dentro de la cadena de mensajes.

A partir de los resultados obtenidos en la tabla II observamos distintos aspectos. En el Foro I, hilo 1 (I, 1), la *palabra clave* «examen» no está en el asunto del hilo del mensaje, lo cual nos revela que «examen» es una palabra importante en el cuerpo del mensaje. La presencia de los meses en los mensajes completos de I, 3; II, 2; III, 1, 3, 4 nos informa del mes al que pertenece ese hilo y, además, que no sólo forma parte de los «créditos» del mensaje, sino que incluso aparece mencionado alguna vez más en el cuerpo del mensaje. Otras cadenas como el Foro II, hilo 4, y el Foro III, hilo 3, no reflejan las mismas palabras clave en el «cuerpo» del mensaje que en el «asunto» sino que la diferencia de palabras clave obtenida del cuerpo se identifica con la propuesta mayoritaria de los coordinadores ante un «asunto» nuevo dentro del mismo hilo. En otros foros podemos encontrar que el «asunto» de una respuesta no tiene relación con el tema del «cuerpo» del mensaje, sino que sólo hay una estructura encadenada. No es éste el caso de los foros de los coordinadores, quienes responden de forma coherente y bien hilada, sin cambiar apenas el asunto. Sólo nueve asuntos se han cambiado en la totalidad de los tres foros (ALY: n.º 76, AFG: n.º 163, 294, 300, 308 y 309; UATD: n.º 378 y JAS: 510; AFVC: n.º 964) para añadir algo nuevo, aclaratorio, creativo u olvidado en el mensaje anterior. Todos los demás mensajes están cohesionados entre sí visual y temáticamente.

Gracias a esta coherencia temática de los mensajes no se produce ningún «discurso1/f» o pérdida progresiva de la cohesión semántica a lo largo de la comunicación. Sin embargo, ejemplificamos en la figura 6 un único mensaje distintivo que rompe con la *Netiquette* de la cortesía, al mismo tiempo que se desvincula del asunto: el mensaje n.º 3, encadenado al primer mensaje del primer foro. Este mensaje n.º 3 está encadenado al n.º 2, pero su cuerpo o contenido no tiene relación con el asunto que lo encabeza. En el mensaje n.º 3 la bienvenida se elude y el «asunto» encierra un cuerpo con *pérdida de coherencia* (Dong, 2004). En este caso debería haberse cambiado el «asunto» o bien dar la bienvenida primero y luego responder al mensaje n.º 2. Esta incoherencia entre asunto y cuerpo es producto de una comunicación asincrónica al tiempo que precipitada. Si procesáramos cualitativamente los cuerpos de estos asuntos de bienvenida, no hallaríamos coherencia alguna. Al igual que forman parte

Tabla II. Palabras clave de hilos de discusión

Foros	*Hilos*	*N.º de mensajes*	*Asunto del hilo de mensaje*	***Palabras clave*** *[frecuencia]* ***(Asunto + cuerpo)***	***Palabras clave*** *[frecuencia]* ***(Sólo cuerpo)***
I Foro 2003-2004	#1	11	«Comité organización jornadas»	comité [17] organización [13] jornadas [14] ponencia [9] examen [6]	datos [11] ponencia [6] examen [9]
	#2	8	«Plantilla página de bienvenida»	plantilla [15] página [15] bienvenida [16]	bienvenida [8]
	#3	12	«Última reunión coordinación del curso 2003-2004»	última [14] reunión [21] fechas [10] mayo [12] coordinación [15]	junio [9]
	#4	9	«Certificado Coordinación»	certificado [12] junio [9] coordinación [10]	[0] «lo» [8] uso de anafórico en vez de «certificado»
II Foro 2004-2005	#1	10	«Enhorabuena»	enhorabuena [14] felicitaciones [5] foto [5]	enhorabuena [5] felicitaciones [5] foto [5]
	#2	11	«Posible taller-workshop sobre learning design»	learning [13] design [13] taller-workshop [12] julio [10] verano [5]	taller [7] interesante [5] vacaciones [6] verano [5]
	#3	13	«Plantilla de página web»	plantilla [30] web [25] página [24] Javier [17] correo [20] tutorial [6]	plantilla [18] tutorial [6] herramienta [10] correo [18]
	#4	14	«Taller work-shop learning design»	design [16] taller-workshop [15] taller [12] learning [15] inglés [7] portátil [5] interesado [6] asistir [8]	taller [12] inglés [7] portátil [5] interesado [6] asistir [8]
III Foro 2005-2006	#1	14	«Prueba nuevos formularios cambio clave»	cambio [25] claves [19] prueba [14] formularios [16] enero [16] contraseña [7]	cambiar [23] claves [14] contraseñas [16] desafío [5]
	#2	10	«Absentismo de los alumnos y Campus Virtual»	absentismo [19] motivación [8]	absentismo [9] motivación [8] juego [6]
	#3	10	«reunión Navidad 05»	diciembre [10] navidad [12] diciembre [10] reunión [12]	martes [7]
	#4	11	«Aplicación para encuestas»	diciembre [19] aplicación [11] encuestas [12]	asistir [7]

de la cortesía los mensajes de bienvenida al inicio de nuevos foros para saludar y poder presentarse nuevos usuarios a medida que se van incorporando al foro, también es importante mantener la correlación entre asunto y cuerpo, cambiar de asunto o abrir un mensaje nuevo para mantener la coherencia discursiva a lo largo de los foros.

Mensaje n.º 1 Autor A
Asunto: bienvenida
un saludo cordial a las nueve de la mañana
Mensaje n.º 2 Autor A
Asunto: re: bienvenida
¿por qué no figura el nombre del autor en el listado general?
Mensaje n.º 3 Autor B
Asunto: re: bienvenida
yo sí que veo el nombre del autor en la página principal del foro

Figura 6. «Asunto» y «cuerpo» incoherente

La aplicación de *Concord*: contexto de las palabras

La herramienta *Concord* nos permite hallar las combinaciones de unas palabras con otras, su frecuencia y su distribución en el discurso. Como aplicación de esta herramienta, nuestro estudio se ha centrado en conocer cuál es el índice de frecuencia de elementos referenciales o anafóricos (Brown & Yule, 1983; Gabriel de Ávila, 2004) y cómo se distribuyen en el foro. Por ello, cuando creamos la palabra prototipo «anaforicospfg» para indexarlos, computamos con *Wordlist* una abundante frecuencia de anafóricos [797]: alguno [19], ello [39], eso [53], lo [592], llegando a considerar que la presencia de anafóricos dificulta el cómputo exacto de *palabras clave*. Sabemos que los elementos pronominales anafóricos o referenciales hacen alusión y sustituyen a palabras o conceptos que han sido nombrados previamente en alguna parte del texto. A pesar de que tales palabras o conceptos ya no aparezcan de forma explícita a lo largo del texto, sin embargo, los sobrentendemos cuando han sido sustituidos por estos elementos pronominales o anafóricos. Un ejemplo evidente de uso de anafóricos en detrimento del cómputo de la *palabra clave* es el Foro I, hilo 4 (véase tabla II). En este mensaje la presencia del anafórico «lo [8]» en el «cuerpo» del mensaje sustituye a la palabra clave «certificado», palabra que no se computa como *palabra clave* en el «asunto» por estar sustituida por «lo».

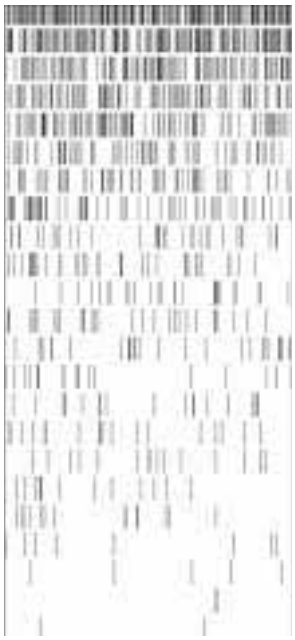
En consecuencia, para computar la recurrencia de los anafóricos en el discurso de todos los foros, sometimos el *corpus* a un análisis más específico. Al analizar con *Concord* las colocaciones de algunas formas anafóricas, obtuvimos los datos de la tabla III. La abundancia de anafóricos en el discurso se muestra en la acumulación de líneas verticales que reflejan la mayor o menor confluencia y situación de estos elementos anafóricos en el conjunto de los foros.

Un estudio más detallado del contexto que nos permitiese identificar cada elemento anafórico con su referente nos aportaría información de qué palabras no llegan a ser computadas como clave porque pierden su capacidad de distinción al ser sustituidas por dicho elemento anafórico.

EVALUACIÓN

Respecto a la fiabilidad y validez para cuantificar términos, *Wordsmith* funciona satisfactoriamente. Un ejemplo significativo para probar su fiabilidad de computación es el término «didácticos» que, a pesar del carácter docente del foro, tan sólo ha aparecido una vez. Cuando hemos comprobado si realmente este término sólo aparecía una vez en los foros, hemos recurrido a la herramienta de «Buscar» de la plataforma WebCT y hemos seleccionado los criterios apropiados para obtener que, efectivamente, el mensaje n.º 1033 es el único que contiene la palabra «didácticos» dentro del foro «Principal». Previamente, el buscador nos informaba que dicho término aparecía en tres mensajes. Ello es debido a que el término se repite en una misma cadena de mensajes al ser el mensaje 1033 «citado», y copiado su texto dos veces. La cita de los mensajes anteriores en un mensaje nuevo supone un incremento de la frecuencia de las palabras en el cómputo, pudien-

Tabla III. Frecuencia de elementos anafóricos en el desarrollo del discurso

Formas anafóricas	*Categoría*	*Frecuencia*	*Gráfica de aparición en el texto*
los que, las que	Relativo	550	
lo que	Relativo	302	
se lo, se la	Pronombres personales	165	
me lo, me la	Pronombres personales	147	
se los, se las	Pronombres personales	134	
nos la, nos lo	Pronombres personales	72	
me los, me las	Pronombres personales	72	
os lo, os la	Pronombres personales	71	
los más, las más	Pronombres personales	39	
os los, os las	Pronombres personales	38	
-arlo,-arla	Infinitivo + Pronombre personal	32	
nos los, nos las	Pronombres personales	30	
algo que	Indefinido + Relativo	28	
te lo, te la	Pronombres personales	24	
-arlos, -arlas	Infinitivo + Pronombre personal	22	
eso que	Demostrativo + Relativo	19	
por eso	Conjunción + Demostrativo	16	
te los, te las	Pronombres personales	15	
por ello	Conjunción + Pronombre personal	14	
aquellos que, aquellas que	Demostrativo + Relativo	8	
unos cuantos, unas cuantas	Indefinidos	5	
-ndolas, -ndolos	Gerundio + Pronombre	2	
-ndola, -ndolo	Gerundio + Pronombre	2	

do falsear los resultado y haciendo preciso un preprocesado de los textos.

Figura 7. Búsqueda realizada del mensaje que contiene «didácticos»

Las interpretaciones temáticas hechas a partir del listado del vocabulario y su índice de frecuencias se han validado mediante la lectura manual de las palabras en su contexto. La herramienta *Concord* se muestra como un método complementario para conocer el contenido del mensaje, ya que permite localizar las palabras de interés en su propio contexto.

También podemos apuntar algunas de las dificultades que hemos tenido para analizar los textos con precisión y sencillez. Desde un punto de vista ortográfico, hemos tenido dificultades para procesar palabras sin tildes, errores tipográficos o palabras amalgamadas. Morfológicamente hemos echado en falta la disponibilidad de un lemario gramatical español que agrupara las formas verbales y los sustantivos y adjetivos, independientemente de su número y género. En el aspecto semántico hemos necesitado crear de forma minuciosa un lemario específico de términos académicos, con sus sinónimos e invocaciones para computar con rigor el máximo de palabras.

CONCLUSIONES

La utilización de una herramienta automática cuantitativa como *Wordsmith* nos ha permitido computar el léxico de unos foros virtuales generados en un ámbito de enseñanza virtual sin leerlos previamente. A partir de la cuantificación del léxico en dichos entornos educativos hemos inferido información cualitativa acerca de la temática de los foros. Los participantes debaten sobre demandas y pres-

taciones técnico-administrativas de la plataforma para el personal docente y discente de la UCM. Además, esta herramienta nos ha permitido obtener información acerca de la correlación léxica, semántica e ideológica a lo largo de los textos. Esto se ha realizado mediante un proceso deductivo basado en los resultados del análisis automático de los mensajes, sus asuntos y sus cuerpos, sin previo conocimiento ni lectura de los foros. En lo que respecta al tipo de discurso de los foros, hemos distinguido cuantitativa y cualitativamente un tipo de discurso personal-emotivo-lúdico, frente a un discurso más académico y formal. Esta diferencia la ha marcado un vocabulario emotivo, expresivo y coloquial que, aunque con baja frecuencia de aparición en el conjunto del *corpus,* destaca por su singularidad.

A través de la identificación automática de *la palabra clave* se ha comprobado que, dentro del foro formal analizado, los mensajes incluidos en hilos de discusión han mantenido cierta coherencia temática elevada, sin caer en una pérdida progresiva de la correlación proporcional a la distancia entre mensajes (discurso 1/*f)*. Lo ideal es que cada mensaje, aunque esté dentro de una cadena, cambie el asunto propio cuando introduzca una idea nueva. El usuario-emisor no sólo reflexionaría sobre las palabras significativas del «asunto» del mensaje, sino que luego las palabras clave serían fácilmente computables e informativas para el receptor-lector. Hemos ratificado con la aplicación de *Keywords* que las *palabras clave* de los mensajes procesados se calculan rápidamente en un análisis conjunto de asunto y cuerpo. En cambio, para obtener las palabras clave del cuerpo del mensaje se requiere un preprocesado más minucioso que incluya las expresiones anafóricas. No obstante, hemos comprobado que el análisis sólo del cuerpo de la cadena de mensajes en muchos casos nos da la respuesta del asunto. Aunque una de las dificultades de no hallar las *palabras clave* en el cuerpo del mensaje es la existencia de sinónimos, no asociados previamente a la *palabra clave,* o de anafóricos no identificados, y que la herramienta *Wordsmith* no puede detectar si no hay un procesamiento previo de estos elementos. Efectivamente, un proceso de asociaciones de sinonimia, anáforas, hiperonimias, etc., para su identificación con palabras clave y la extracción de su significado requiere la intervención previa del humano, el empleo de herramientas más potentes (Plan Nacional de Investigación: Vol. II: Áreas prioritarias: 409-419) o modelos de computación lingüístico-matemática más complejos (LSA: Latent Semantic Analysis) que están actualmente en pleno desarrollo.

También hemos apreciado que los profesores tienen un *conocimiento compartido* en los foros. Al igual que las conversaciones, éstos tienen un discurso referencial y alusivo por excelencia. Se alude a palabras, contextos o conceptos mencionadas o conocidos con anterioridad por todos, pero que no aparecen en el texto. Por tanto, en muchos mensajes no se identifican ni computan como *palabras clave* ya que no están explícitas sino que se las alude con elementos anafóricos o catafóricos. De ahí que hayamos procedido a identificar un pequeño conjunto de palabras que hacen referencia a una palabra clave o señalan un concepto mencionado con anterioridad.

En una fase posterior de este estudio se pretende aplicar las herramientas de la lingüística computacional a la evaluación automática de foros de debate extensos generados por numerosos participantes. De esta manera, se podrá validar un método que evalúe no sólo el índice de participación de los estudiantes, sino la calidad de las aportaciones. Además, el profesor podría procesar la información desde una perspectiva inversa a la nuestra, ya que participaría activamente fomentando la generación de los textos. Estos textos podrían emplearse para la creación de *corpus* específicos de la materia impartida. Para el profesor los foros se convertirían en un registro cuantificable del proceso de aprendizaje de los estudiantes.

BIBLIOGRAFÍA

ÁVILA OTERO, G. (2004): «Algumas Consideraçoes sobre a Importancia da Continuidade Topica na Classificaçao Automática de Documentos Digitais». http://www.geocities.com/gabriel_othero/public_arquivos/continuidade_topica.pdf (acceso, mayo 2006).

BROWN, G., y YULE, G. (1983): «The nature of reference in text and in discourse», *Discourse analysis*. Cambridge: Cambridge University Press, pp. 191 y ss.

DAVIES, Mark: *Corpus del español*. http://www.corpusdelespanol.org (acceso, mayo 2006).

DAVIES, M. (2006): Comunicación personal.

DONG, A. (2004): «Quantifying coherent thinking in design: a computational linguistics approach», *Design Computing and Cognition 2004 (DCC04)*, J. S. Gero (ed.), Dordrecht: Kluwer Academic Publishers, pp. 521-540.

FERNÁNDEZ-PAMPILLÓN, A. (2005): «Herramientas de análisis textual: Claves para el análisis de la "información" textual», *Las nuevas profesiones de las lenguas*. Editora Miriam Llamas Ubieto, Madrid, Liceus, pp. 79-90.

FERRERO, P.; ALDA, J. (2005): «La tutorización virtual y la expresión de las emociones», Actas II Jornada Campus Virtual UCM, pp. 129-133.

GRAESSER, A. C.; MILLIS, K. K., y ZWAAN, R. A. (1997): «Discourse comprehension», *Annual Review of Psychology,* 48, pp. 163-189.

JUST, M. A., y CARPENTER, P. A. (1987): «Syntactic Structures and Syntactic Processing». *The Psychology of Reading and Language Comprehension*. Allyn and Bacon Inc, pp. 138 y ss.

LANDAUER, T. K. (1999): «Latent semantic analysis: A theory of the psychology of language and mind», *Discourse Processes* 27(3), páginas 303-310.

Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica. Vols. I, II, III. Comisión Internacional de Ciencia y Tecnología. Ministerio de Ciencia y Tecnología. 2004-2007. http://www.madrimasd.org/queesmadrimasd/pricit/documentos/dfault.asp?id=41&IdDoc=2301&op=O&detail= (acceso, mayo 2006).

PRESSLEY y AFFLERBACH (1995): «What readers Can Do When They Read: A Summary of the Results From the on-Line self report studies of Reading, Identifying and Learning Text Content», *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. Lawrence Erlbaum Associates, Inc., pp. 30 y ss.