INVENTARIO DE PALABRAS CLAVE TEMÁTICAS PARA LA CLASIFICACIÓN AUTOMÁTICA DE NOTICIAS DE TELEVISIÓN

Ángel Rodríguez Bravo*

Dpto. Comunicación Audiovisual i Publicidad II. Facultat de Ciències de la Comunicació. Universitat Autònoma de Barcelona.

Lluís Mas Manchón**

Dpto. Comunicación Audiovisual i Publicidad II. Facultat de Ciències de la Comunicació. Universitat Autònoma de Barcelona.

Resumen: En el marco de un proyecto financiado por el CAC (Consell de l'Audiovisual de Catalunya), se realizó una aproximación comunicológica al problema de la selección de palabras clave para la clasificación temática de noticias de televisión a partir de sistemas de reconocimiento automático. Aplicamos análisis del discurso (entorno al concepto "tema"), teoría de la noticia y técnicas *lexicométricas* y de recuperación de la información, para definir un Protocolo Integral de Selección de Palabras clave. Del trabajo de 4 investigadores con este protocolo sobre una muestra transcrita de 698 noticias ha resultado un *lexicon* de 1000 palabras clave distribuidas en 15 temas, contrastado mediante el estadístico Lambda de Wilks.
Palabras clave: Información audiovisual; clasificación temática; palabras clave; temas de noticias televisivas; lexicometría; sistema de clasificación.

Title: LEXICON OF THEMATIC KEYWORDS FOR THE AUTOMATIC CLASSIFICATION OF TV NEWS.
Abstract: In the framework of a research project funded by CAC, a communication approach was taken to the problem of keywords selection for the themes indexing of TV news by word spotting. This is, we apply discourse theories (concept of "themes"), news theory and lexicometry and information retrieval techniques, for the definition of a complex Protocol of Keywords Selection. The work of 4 researchers with this protocol on a 698 transcript news sample resulted in a lexicon of 1000 keywords distributed in 15 themes, which is contrasted statistically with Lambda of Wilks.
Keywords: Audiovisual Information; themes indexing; keywords; themes in TV news; lexicometry; retrieval system.

* Angel.Rodriguez@uab.cat
** lluis.mas.manchon@gmail.com

Recibido: 25/03/2011; 2ª versión: 16/05/2011; aceptado: 27/05/2011.

RODRÍGUEZ BRAVO, A. y MAS MANCHÓN, LL. Inventario de palabras clave temáticas para la clasificación automática de noticias de televisión. *Anales de Documentación*, 2011, vol. 14, nº 2. Disponible en: <http://revistas.um.es/analesdoc/article/view/123271>.

1. INTRODUCCIÓN

En el año 2004, el *Consell de l'Audiovisual de Catalunya* (CAC)[1] encargó una investigación al Laboratorio de Análisis Instrumental de la Comunicación (LAICOM)[2] para iniciar el trabajo sobre el diseño de una herramienta automática de segmentación y clasificación temática de noticias en televisión. El objetivo final era desarrollar un sistema que procesara los informativos en tiempo real, segmentándolos en unidades-noticia y que fuese capaz, además, de clasificar temáticamente cada uno de estos segmentos. En función del carácter esencialmente textual de la información televisiva y del grado de desarrollo sobre el reconocimiento automático de formas en el ámbito del habla, decidimos abordar el problema a partir del análisis del audio de los informativos, concretamente del estudio de la locución. Esta investigación se dividió en dos fases: a) formalización de los criterios básicos que rigen la clasificación temática de noticias de TV y, b) localización de patrones prosódicos destinados a orientar la segmentación automática del informativo en las diferentes noticias que lo componen. En este artículo presentamos los resultados de la primera etapa, es decir, la investigación dirigida a localizar un inventario específico de palabras clave, cuyo objetivo último es automatizar la clasificación temática de noticias.

En los últimos años, desde el campo de la ingeniería se han desarrollado con gran éxito algoritmos de reconocimiento de palabras (Abberley *et al*., 2006; Nakamura *et al*. y otros), lo que ha abierto grandes oportunidades para la configuración de inventarios de palabras cuya presencia en una noticia indique el tema dominante de la misma. Estos tipos de sistemas son denominados *word spotting* y no han dejado de mejorar gracias a los avances en los algoritmos matemáticos de reconocimiento de patrones y las herramientas de predicción estadística, como los Modelos de Markov. Así, en tanto que disponíamos ya de sistemas razonablemente eficientes para el reconocimiento sonoro de palabras concretas, si en esta investigación lográbamos descubrir cuales son los conjuntos finitos de palabras clave semánticamente dominantes para cada uno de los respectivos temas noticiosos, tendríamos en nuestras manos un inventario de palabras capaz de clasificar temáticamente cualquier noticia.

A pesar del interés evidente de este tipo de estudios, desde el campo de la comunicación no se han investigado protocolos de selección de palabras asociadas a los temas noticiosos. De hecho, hasta ahora han sido casi siempre los ingenieros que desarrollan los sistemas de reconocimiento los que han propuesto los modelos (normalmente probabilísticos) para la selección de palabras clave, tomando como único criterio el número de apariciones de las palabras en una muestra estratificada (procedimientos lexicométricos). Este criterio tiene muchas limitaciones, algunas de las cuales son evidentes y pueden ser solventadas con protocolos muy sencillos (está claro, por ejemplo, que las preposiciones, los pronombres o los conectores no pueden ser palabras clave temáticas), pero otros problemas no son tan evidentes y sin embargo son centrales para mejorar la eficiencia de tal sistema de clasificación.

En nuestra investigación, hemos entrado más a fondo en las implicaciones teóricas de la selección de palabras vinculadas a un campo semántico determinado para aclarar a qué tipo de limitaciones nos referimos. Tomando el discurso como estructura integral, jerárquica y solidaria (van Dijk, 1990), es evidente que nos interesa el nivel pragmático (campos temáticos percibidos como tales a partir del nivel léxico del "discurso noticia". Este abordaje holístico mostrará las limitaciones de aplicar un criterio exclusivamente estadístico y señala, a la vez, la necesidad de partir de un enfoque semántico. Básicamente, a nivel pragmático, la noticia como discurso asociado a un tema no es una entidad cerrada y hermética. El criterio temático pone en liza distintos campos semánticos que se superponen; pensemos, por ejemplo, en las secciones informativas definidas por los conceptos: "Sociedad", "Conflicto Social" y "Sucesos", y en el modo en que éstas se solapan con el criterio geográfico que define la sección que suele denominarse "Internacional". Si a esto añadimos que toda noticia tiene un alcance geográfico y temporal que determina el tipo de palabras utilizadas, y que, además, el formato de cada programa y el estilo de cada periodista condicionan globalmente la terminología de una noticia, es evidente que utilizar solamente criterios estadísticos para la localización de palabras clave es ineficiente.

Indiscutiblemente, el género informativo tiene un lenguaje propio que determina el uso de unas palabras u otras, en función de la parte de la realidad que se quiere referenciar (nivel semántico) y en función de un tipo de redacción y dicción preestablecidos (nivel léxico y morfológico). Consideramos, en consecuencia, que un abordaje comunicológico del problema nos permitiría ponderar de forma integral los diferentes criterios que dan valor temático a la palabra, y así fundamentaremos nuestra afirmación de que el criterio *lexicométrico* debe someterse a los criterios funcionales y semánticos de la palabra en el discurso, y no a la inversa[3].

2. METODOLOGÍA

Como el lector habrá observado ya en la Introducción de este artículo, el problema metodológico al que nos enfrentamos se centra en dos premisas:

1. Estudiar el modelo de clasificación temática de noticias utilizado por el servicio de análisis de contenidos del CAC (*Consell de l'Audiovisual de Catalunya*), y evaluar sus posibilidades para ser implementado en un sistema automático
2. Estudiar el modelo del sistema de reconocimiento automático del habla *word spotting*

Estas dos premisas han determinado una metodología que estudia el criterio de un periodista para redactar una noticia de un determinado tema usando unas palabras, y el criterio de un analista de contenidos o un documentalista cuando elige palabras como términos de búsqueda temática de noticias.

Obviamente, dadas estas condiciones de trabajo iniciales, nuestro punto de partida para definir y acotar cada campo temático fueron las noticias del archivo del CAC, si bien bajo una hipótesis de investigación para poder generalizar los resultados a universos infinitos:

La decisión de clasificar una noticia en un campo temático determinado, depende del reconocimiento de un conjunto finito de palabras clave que tienden a quedar asociadas a este campo por la práctica profesional periodística. Este conjunto de palabras clave es específico de cada campo y varía en un porcentaje muy alto entre un campo temático y otro.

Definida la hipótesis, nuestro objetivo era muy claro: localizar estos grupos de palabras mediante un procedimiento que tuviera en cuenta todos los criterios, y, a continuación, probar la bondad estadística de estas palabras sobre una muestra. Por tanto, estas son las fases de la metodología seguida:

1. Revisión del estado del Arte

1.1. Investigación documental:

- Análisis del discurso y estructura de la noticia
- Criterios de clasificación informativa en campos, secciones, temas, etc.
- Biblioteconomía
- Lexicometría
- Sistemas automáticos de selección de contenidos
- Reconocimiento automático de voz (sistemas *word spotting*)

2. Preparación de un inventario eficiente de palabras clave

2.1. Selección de una muestra significativa de noticias:

- Muestra aleatoria estratificada de 698 noticias procedentes de las emisiones de informativos televisivos realizadas a lo largo de un año (1/07/2004 - 30/06/2005) por los canales: TV3, Canal33, TVE, BTV, CityTV y Localia.

2.2. Fundamentación y primera construcción del inventario hipótesis:

- Se parte del criterio de clasificación de noticias utilizado por el CAC, que en ese momento está organizado y definido en las 15 categorías siguientes: *art i cultura, ciència i tecnologia, conflicte social, crònica internacional, crònica política, economia i negocis, educació i ensenyament, esports, medi ambient, mitjans de comunicació, sanitat, societat, treball, trànsit* y *temps*, más un campo complementario de "*altres*".
- Se diseña un procedimiento de análisis de la muestra de noticias para construir un inventario hipotético constituido por 15 conjuntos de palabras clave asociadas respectivamente a cada una de las 15 categorías en las cuales el CAC clasifica cualquier noticia. El método de análisis diseñado se estructura en dos fases: una subjetiva y una intersubjetiva, basadas en el conocimiento experto de 4 investigadores.

2.3. Verificación y depuración iterativa del inventario-hipótesis:

- Test de eficacia-1: Una vez hecha la primera construcción del inventario hipotético siguiendo los procedimientos subjetivo e intersubjetivo, se realizó un primer test de eficacia, con el objeto de comprobar con que efectividad este inventario clasificaba una selección aleatoria de 20 noticias.
- Depuración y corrección: Realizado el análisis de los resultados se procedió a depurar y complementar el inventario para mejorar su capacidad de clasificación temática en función de las deficiencias detectadas en el test.

- Test de eficacia-2: Hecha la mejora del inventario se procedió a realizar un segundo test de eficacia, comprobando ahora con qué efectividad quedaban clasificadas por nuestro inventario 45 noticias más (3 de cada categoría temática) seleccionadas aleatoriamente.
- Depuración y corrección con procedimientos lexicométricos: Realizado el análisis de los resultados del segundo test, se procedió a depurar y complementar por segunda vez el inventario para mejorar su capacidad clasificadora en función de las deficiencias detectadas en el test. Esta segunda complementación se realiza ya desarrollando análisis lexicométricos de frecuencias para localizar nuevas palabras clave no detectadas con el procedimiento anterior. El resultado final de esta tarea fueron las 15 listas de palabras que constituyeron el INVENTARIO-HIPÓTESIS de palabras clave.

3. Contrastación estadística del INVENTARIO-HIPÓTESIS:

- Construcción de un fichero de datos: Cerrado el inventario hipotético, se preparó un fichero con los datos de aparición y repeticiones de todas las palabras del inventario en la muestra completa de noticias.
- Análisis discriminante: Una vez que consideramos satisfactorio el inventario hipotético, se continúa la investigación contrastando estadísticamente la capacidad de cada uno de los 15 grupos de palabras de este inventario para clasificar noticias en sus respectivos campos temáticos. Para hacerlo se utilizaron procedimientos de análisis discriminante.
- Última depuración y selección definitiva de palabras clave: A partir del análisis estadístico se pudieron seleccionar ya solamente las palabras clave de las que se había podido comprobar en el test estadístico que tenían una capacidad discriminante significativa y eficaz. Este último paso nos permitió, finalmente, elaborar el listado definitivo de palabras clave para cada una de las 15 categorías analizadas.

El enfoque metodológico general que hemos dado a esta investigación es el de un estudio comunicológico. Más allá de la tradicional tendencia *descriptivista* de este vasto campo, nosotros abogamos por la *comunicología* como disciplina cuyo objeto de estudio es la función comunicativa de algún aspecto físico del mensaje. La eficacia de esta función se estudia aplicando un Análisis Instrumental que mide y formaliza este aspecto físico y un test de recepción que controla su impacto comunicativo (Rodríguez Bravo, 2003). De este modo, es posible saber con certeza si el mensaje sirve a sus objetivos y provoca unos efectos, o no. Esto implica que las cuestiones semánticas de ese mensaje pueden ser evaluadas. Por tanto, si nuestro objeto de estudio es el "léxico de la noticia que remite a los diferentes temas", podremos decir que:

- El presentador ha elegido una palabra u otra, acompañada de unas palabras u otras, etc. en función de la contribución temática que hace/n
- se puede medir la contribución temática de cada palabra

Para el estudio necesitamos saber qué criterios ha puesto en funcionamiento el periodista para elegir las palabras mientras prepara sus noticias. Las distintas teorías de la noticia nos indican todas las fases por las que pasa una información desde el evento de la que proviene hasta su transmisión en forma de noticia, en función de los ámbitos social,

temporal y geográfico y los condicionantes económicos, ideológicos, demográficos… Por tanto, en primer lugar, debíamos echar mano de aquellos cuerpos de conocimiento que estudian las palabras empleadas en la noticia en función del lenguaje periodístico. Nos referimos principalmente a cuestiones de géneros informativos, temas y campos semánticos.

Ante esto, nos encontramos con dos matices diferenciales (o problemas de conocimiento) de nuestro objeto de estudio que condicionaron todos los pasos de nuestro trabajo:

- Nuestro mensaje se limita al léxico: por tanto, ¿de todas las formas que el emisor tiene de marcar el mensaje físicamente con un tema, cuáles envuelven exclusivamente al léxico?
- Nuestro "receptor" acabará siendo una máquina (nuestros resultados deberán se programados en un sistema automático): ¿qué lenguaje deberemos emplear con esa máquina para transmitirle nuestro mensaje y para que posteriormente "ella" lo maneje?

No todas las palabras tienen función semántica y pragmática, pero sí todos los referentes semánticos textuales se constituyen de palabras. Debíamos tener en cuenta, en consecuencia, la cibernética y el *word spotting* como fenomenologías determinantes, para que los criterios de selección y organización de palabras clave tomasen como punto de partida las condiciones impuestas por ellos.

Con este marco teórico en la mano, dispondríamos de los criterios básicos a tener en cuenta para elegir las palabras clave y estaríamos en condiciones de iniciar el proceso de selección siguiendo los dos principios básicos siguientes:

- Definición de unidades de análisis modélicas: se eligieron casos de noticias para cada tema que resultaban paradigmáticos. Se identificaban los campos semánticos de cada tema y se elegían noticias representativas de cada uno de ellos. En el caso de "Tiempo", por ejemplo, debían haber informaciones típicas de invierno, verano, otoño y primavera.
- Análisis cualitativo, iterativo e intersubjetivo: los criterios de selección de palabras clave fueron probados sucesivamente en esas "unidades modélicas" definidas mediante protocolos intersubjetivos, es decir, asegurando la coincidencia de todos o una mayoría cualificada concreta de investigadores en la elección de palabras.

Con la garantía de que los criterios contemplaban los diferentes campos semánticos de cada tema y que los investigadores tenían en cuenta la función discursiva temática de la palabra, se definieron, luego, unos pasos de análisis sistemáticos de la muestra de noticias en forma de protocolo definitivo. Las palabras clave seleccionadas en el protocolo serían sometidas a análisis lexicométrico mediante el programa Concordance en una muestra de 698 noticias de diferentes temas, que daría como resultado un "Inventario-Hipótesis" de palabras clave. Estas palabras fueron evaluadas, finalmente, mediante el análisis discriminante de la Lambda de Wilks, para estudiar cual era su contribución estadística a

la definición de su tema-grupo. En función de los resultados de ese análisis, se seleccionó la menor cantidad de palabras posible para cada grupo temático.

3. MARCO TEÓRICO: NIVEL LÉXICO-SEMÁNTICO Y ESTRUCTURA TEMÁTICA DE LA NOTICIA

El proceso de redacción de la noticia se caracteriza por una serie de fases, la primera de las cuales corresponde al gatekeeper y consiste en determinar si una información es noticia o no y por qué. O lo que es lo mismo, cuáles son los elementos de novedad-actualidad en el marco de un contexto, los cuales serán traducidos en palabras. Según las características del medio -su alcance, medios técnicos y humanos, su enfoque, etc.-, el gatekeeper podrá elegir un elemento noticioso u otro de un mismo evento. Por ejemplo, partiendo de una eventual ola de frío en primavera, el evento noticioso podrá ser:

- El fenómeno meteorológico
- La cantidad de muertos que deja
- El consumo récord de energía de la región
- La mala gestión política que se hace de la situación o
- El hecho de ocurrir muy cerca del ámbito de emisión del canal

Con lo que en realidad, no hay un único evento noticioso, sino tantos como "aspectos noticiosos" escogidos. Por encima de todos los criterios de *noticiabilidad* (novedad, negatividad, proximidad…), toda noticia es presentada como algo desconocido o nuevo sobre un ámbito conocido, es decir, cada noticia es presentada como novedad ("*news*" en inglés) de un tema particular. De forma que un evento no pertenecerá necesariamente a un único ámbito temático, sino que será específico del particular aspecto noticioso que se resalte. Decimos, por tanto, que cada aspecto noticioso lleva consigo códigos de *tematicidad*. En el ejemplo anterior, podríamos estar hablando de "Tiempo", "Sucesos", "Medio Ambiente", o incluso de "Política".

Ahora bien, cada aspecto noticioso pone en marcha unos campos semánticos de la información que llevan consigo criterios subsidiarios de *tematicidad*. Así, cuando hablamos de la "muerte de dos indigentes en la calle" (tema "Sucesos"), deberemos decir que su muerte se debe al "intenso frío de estos días" (tema "Tiempo"), fenómeno que "las autoridades no han sabido predecir ni gestionar" (tema "Política"). Por tanto, una vez se ha elegido el aspecto noticioso de la información, el gatekeeper o periodista define la jerarquía temática (o los campos semánticos) en los que se enmarca, incluido el "tema" o "sección" general del Informativo. Este proceso determina una dinámica tema/rema (información nueva o temática) que recurre a series de palabras claves de los sucesivos campos temático-semánticos que representan:

"Los temas [en el lenguaje periodístico], por definición, controlan los significados locales, y por ende los posibles significados de la palabra y, por lo tanto, la elección del léxico" (van Dijk, 1990, p. 114).

Un ejemplo muy claro de palabras que enseguida pasan a formar parte de la estructura del discurso periodístico referido a cierto tema son los nombres de países en la categoría

"Internacional", por lo que estas palabras tendrán más valor y serán realmente "claves" al aparecer juntas. Tal y como hemos hecho en el ejemplo del evento de la ola de frío, nuestro protocolo de localización de palabras clave deberá ser capaz de dilucidar palabras asociadas a las categorías temáticas que se configuran en los temas predefinidos por el CAC y, a continuación, estudiar las palabras estructurales de cada una de estas categorías. Por tanto: "por *tematización* entendemos el proceso discursivo mediante el cual un referente se convierte en el asunto principal del discurso" (Brown y Yule, 1993). Ese "referente" será el actor principal y futura palabras clave: "Francia", "armas químicas" o "ministro".

Con la estructura de la noticia ya definida, se echa mano del lenguaje periodístico, para referenciar de forma literal y precisa las realidades del mundo, es decir, para redactar el texto oral del informativo en televisión (Kayser, 1961, 1966, citado en Martínez Albertos, 1998). Los rasgos de esta escritura se sitúan según van Dijk en 5 niveles:

1. Nivel Fonológico
2. Nivel Léxico
3. Nivel Sintáctico
4. Nivel Semántico
5. Nivel Pragmático

Cada uno de los cuales y todos en su conjunto cumplen funciones semánticas. Nosotros queremos estudiar aquí si únicamente con el nivel léxico nos basta para explicar la función semántica de tema. Para ello, partiremos de la función de las palabras y evaluaremos su nivel de manifestación semántica (temática).

Siguiendo con el desarrollo lógico de elaboración de la noticia, sabemos que el trabajo del redactor sería ahora utilizar un lenguaje periodístico en el que prima la elección de palabras capaces de referenciar los campos semánticos definidos de acuerdo a los siguientes criterios:

- *Referencia Lingüística*:
    - Nivel Fonológico: para poder realizar una melodía enfática y muy expresiva durante la enunciación de la noticia en TV, se utilizarán palabras de 3 ó 4 sílabas y con mucha sonoridad –sílabas tónicas largas con vocales "a", "e" y "o", muy reconocibles.
    - Nivel Morfológico (Maniez, 1992 [1987], p. 198): se utilizan verbos en forma activa y en indicativo y palabras llanas y simples en giros directos. Las palabras más informativas suelen ser sustantivos y algún verbo de acción. Los artículos y preposiciones no suelen formar parte de este lenguaje; pero en ocasiones debemos tener en cuenta cierto uso de los mismos. Por ejemplo, la forma de la palabra periodística referencial "Casablanca" (ciudad de Marruecos) respecto de "La Casablanca" (casa del presidente de EEUU), o la de "Madrid" respecto de la de "el Madrid". Esto nos revela la conveniencia de evaluar, también, en esta misma línea otros fenómenos morfológicos como: homofonías (okupa y ocupa), uso del singular o plural, siglas, palabras compuestas, etc.

  - Nivel Sintáctico: al ser un lenguaje de producción colectiva y recepción masiva (Martínez Albertos, 1998), combina la "corrección" de la lengua culta con alguna nota coloquial de alto valor informativo. Es además un lenguaje conciso y claro -predominio de sintagmas nominales en frases cortas (30-36 sílabas/frase, 15-17 palabras/frase). Siendo más concretos, se produce cierta organización sintáctica de las palabras que generan conceptos típicamente periodísticos; por ejemplo, la expresión "de 0 a 3 años" no remite necesariamente al campo "*Educació i ensenyament*", pero su uso sintáctico para expresar una realidad educativa potencia su orientación hacia el tema "*Educació i ensenyament*".
- *Referencia Semántica:* el redactor combina y hace coincidir tres fuentes de significación de la palabra:
  - Términos literales: el diccionario nunca deja de ser la fuente primera en la que se originan los posibles usos semánticos de las palabras. A cada campo semántico le corresponde un conjunto finito de palabras. Se debe controlar que una misma palabra no represente diferentes campos semánticos; por ejemplo: el 'paro' como desempleo y el 'paro' como parada de trabajo.
  - Términos de uso: el redactor comprobará que el significado más común de la palabra se refiere al tema y a los referentes del tema que está referenciando. El redactor selecciona las palabras principalmente por su contribución a la comprensión del hecho por parte de la audiencia, por lo que utilizará el significado socialmente más extendido. No podemos olvidar que muchos temas de noticias se refieren a realidades cotidianas y que reflejan el día a día de la sociedad. Se trata de palabras cuyo hábitat es la calle y son muy eficaces para referenciar o comunicar realidades complejas: "okupas", "mileuristas", "píldora del día después", etc.
  - Términos asociados: en base al sentido que diferentes términos dan al tema, Maniez (1992, p. 215) propone 4 tipos de conexiones semánticas entre palabras referenciales:
    - Implicación total (sinonimia): la inclusión de sinónimos dará fuerza (redundancia) a la categoría.
    - Implicación fuerte (especificidad): una palabra puede derivarse de otra y hacer las funciones de la anterior, que por alguna razón no conviene como palabras clave. Por ejemplo, el ámbito "terrorismo" pertenece al tema "Nacional", pero puede aparecer en "Crónica internacional" también, por lo que se crean palabras de los tipos de terrorismo: "etarras", por ejemplo. Otro ejemplo son las palabras asociadas a "guardia": "urbanos", "nacionales", "guardias civiles", "*mossos*"…
    - Implicación media (asociativa): puede que cierta palabra aparezca siempre simultáneamente a otra, aunque no tengan una relación de sinonimia, permitiéndonos incluir nuevos

ámbitos del tema. Así por ejemplo, podría darse el caso de que siempre que aparece "televisión" (palabras clave de la categoría *"Mitjans de comunicació"*), también aparece "espectador", la cual es a la vez palabras clave representativa de ámbitos artísticos como el "cine" o el "teatro", haciendo así más eficiente el inventario.

- Implicación leve (compatibilidad): es el mismo caso anterior, pero con una relación entre palabras más lejanas (por ejemplo, si el tema es el "Terrorismo", una palabras clave podría ser "coche bomba"). Otro ejemplo es "cocaína", que no es determinante de ningún tema en exclusiva, pero lo es mucho más de la categoría *"Conflicte social"*, que de *"Societat"*, por lo que podría contribuir a diferenciar ambas categorías.

- *Referencia contextual* (social, cultural, temporal): ya hemos dicho que la clasificación temática del CAC es un constructo que aglutina y solapa campos semánticos de diferente índole. Aún así, a cada tema le corresponde cierto tipo específico de palabras:
  - Economía: datos, porcentajes y cantidades en general
  - Política: nombres de protagonistas, verbos de acción y fechas
  - Urgente: cuándo, cómo y por qué
  - Conflicto Social: quién y cómo
  - Internacional: países, gentilicios, etc.
  - Sanidad/Deportes/Tránsito/Trabajo/Educación/Medios/Ciencia: vocabulario específico de los diferentes campos, y en general: lenguaje más coloquial o informal

  Por otra parte, la referencia a ciertos personajes, fenómenos o conceptos, también dependerá de su estabilidad contextual. En este sentido, el lenguaje periodístico tiende a ser conservador y utilizar los nombres más genéricos. Por eso sólo se consideran como parte de este lenguaje algunos nombres de países y personalidades. En general, los nombres de los presidentes de Estado son efímeros y desaparecen de la escena pública repentinamente, y por eso vienen normalmente asociados a vocabulario genérico del tipo "presidente" o "jefe de estado", que son las verdaderas palabras clave. En contraste, los grandes artistas son inherentes a los ámbitos temáticos ("Picasso", "Mozart", etc.) y por eso son palabras que aparecen por sí mismas y son claves del campo en cuestión.
- *Referencia mediática*: la comunidad periodística ha acordado generalmente utilizar el significado de la palabra para referirse a cierto hecho, tema, lugar, persona…, ya que los hechos o fenómenos complejos son denominados con palabras poco comunes y difusas para la audiencia. Los periodistas adoptan expresiones, eufemismos, pleonasmos, adjetivos análogos, etc. (Charaudeau, 2003), que se repiten en todos los medios y programas, y se consigue institucionalizar la forma de decir de esos fenómenos. Estas palabras también serán consideradas como criterio semántico clave: "la guerra del agua", "Euribor", "batalla política por el control de", "atascos quilométricos", "el titular

de educación", "su homólogo checo/italiano", "desplome bursátil", "pisos patera", "contraataque", "compañera sentimental", "inmigrantes ilegales", y un largo etcétera. La mayoría son palabras que habitan en los informativos, los libros de estilos y poco a poco en la calle. Pero hay unas pocas palabras con función conectiva que son propias del lenguaje periodístico: "con el reportero", "con exteriores", "con el Senado", "con el Sánchez Pizjuán", "con París", "con el lugar de los hechos", "según fuentes cercanas/oficiales", etc.

La estabilidad de las funciones referenciales de las palabras vendrá dada por el género periodístico al que pertenece el mensaje informativo: género informativo referencial (Cebrián Herreros, 1994). Ante la actual vorágine informativa y la disposición de medios técnicos, los criterios para la información, persuasión y entretenimiento de las noticias se limitan a la consideración noticiosa de unos campos semánticos u otros, por lo que se mantendrá la estabilidad formal y de género. Ese objetivo marco garantizará el proceso estructural de selección de palabras temáticas que hemos descrito.

Teniendo en cuenta estas 4 capacidades referenciales temáticas del nivel léxico en la noticia (lingüística, semántica, contextual y mediática), exponemos a continuación con mayor detalle los distintos pasos para la construcción del protocolo de selección de palabras clave que las utilizó fielmente como criterios centrales de trabajo.

4. PREPARACIÓN DE UN INVENTARIO DE PALABRAS CLAVE

Hasta aquí hemos revisado los aspectos de lenguaje periodístico que se ponen en funcionamiento para la elección de palabras en función del tema al que pertenecen. En este apartado exponemos ahora todo el procedimiento que siguió el equipo de investigación para implementar los criterios expuestos más arriba en la selección de palabras clave analizando una muestra de noticias. Este protocolo se rige por las dos fases generales del análisis instrumental (Rodríguez Bravo, 2003):

1. *Estudio cualitativo preliminar*: aproximación al vocabulario temático de los periódicos, ya que se considera que la prensa escrita se rige por una clasificación temática más ajustada y utiliza un vocabulario muy exacto para explicar los fenómenos que ocurren cada día (Mizoguchi, 1998).
2. *Selección de una muestra significativa de noticias*: como condiciones muestrales teóricas, definimos un universo de 47.054 noticias, con un error asumido de 4%, una confianza del 95.5% y una varianza de 50.50 (Sierra Bravo, 2001), lo que exigía que la muestra fuese de al menos 551 noticias, estratificada en temas y canales. También tuvimos en cuenta que la cronología de la muestra debía representar los ciclos anuales para incluir todo el rango de eventos que ocurren anualmente (Donough y Gish, 1995). A partir de aquí, se diseñaron las muestras para las diferentes fases del protocolo:
    1. *Primer desarrollo del "Inventario-hipótesis de palabras clave"*: fase iterativa que se realizó con pequeñas muestras de 5 noticias de cada tema, analizadas por cada uno de los investigadores, hasta contabilizar un total de 84 noticias.

2. *Análisis Extensivo*: el análisis discriminante del *"Inventario-hipótesis de palabras clave"* se realizó sobre una muestra de 698 noticias.

3. *Fundamentación y primera construcción del inventario-hipótesis*: este paso se refiere directamente al proceso de selección de palabras. Esta etapa tenía ya el objetivo específico de seleccionar un conjunto de palabras clave para cada de las 15 categorías temáticas predefinidas. El procedimiento de partida fue seguir los criterios que propone Pinto (1997). Se definieron 5 pasos de control del procedimiento de selección hasta llegar a una primera selección de palabras clave:
    1. *Validación de la metodología subjetiva*: se comparó la elección y contabilización manual y *lexicométrica*. La contabilización era sistemáticamente idéntica mediante ambos procedimientos, aunque en la selección lexicométrica aparecían con frecuencias muy altas numerosas palabras semánticamente irrelevantes para el campo temático: artículos pronombres, verbos transitivos, adverbios temporales, etc.
    2. *Metodología subjetiva*: puesta en común del criterio de selección de palabras clave de los investigadores mediante el análisis de 6 noticias según los criterios generales:
        a. Identificación del campo temático de la noticia y búsqueda introspectiva de palabras teóricamente conectadas a ese campo.
        b. Localización y selección de las palabras que efectivamente aparecían en la noticia.
        c. Selección de las palabras que aparecían y que teóricamente actuarían como patrones de búsqueda de un documentalista.
    3. *Metodología ínter-subjetiva*: las palabras seleccionadas mediante el método subjetivo fueron revisadas utilizando un método ínter-subjetivo, que consistía en una puesta común de los investigadores siguiendo el camino inverso al procedimiento anterior. Con ello, pudimos medir el porcentaje de coincidencia entre investigadores en la selección de palabras. En la "Tabla I" puede verse el grado de coincidencia intersubjetiva de los investigadores para la selección de palabras clave (87,31% de coincidencia y 9.74 de desviación estándar). Este estudio reflejaba la coherencia del trabajo realizado y apoyaba la fiabilidad de esta metodología. Lográndose, además, formular un procedimiento general de selección de palabras clave a partir de los siguientes criterios:

	Not-1	Not-2	Not-3	Not-4	Not-5	Not-6	Coinc. Mitja	Desv. Estand.
Art	85	93,1	91,6	84,3	63,46	92,5	84,99	11,22
Cie	96,6	88,8	96,6	81,4	66,7	66,6	82,78	13,71
Cons	100	87,5	75	75	91,67	77,78	84,49	10,26
Int	94,7	77,08	88,8	75	86,36	100	86,99	9,74
Cro	100	100	100	94,4	83,3	100	96,28	6,74
Eco	77,9	90	100	100	91,6	72,3	88,63	11,41
Edu	100	91,6	87,5	72,2	90	96,4	89,62	9,65
Esp	88,8	90,48	95,46	91,1	95,2	94,4	92,57	2,81
Med	62,9	81,2	100	87,5	68,7		80,06	14,81
Mit	100	100	100	88,8	94,4	83,3	94,42	7,05
San	77,7	83,3	93,7	85	69,2		81,78	9,08
Soc	78,5	91,6	68,7	75	85	93,7	82,08	9,76
Tre	85	75	100	91,6	75	100	87,77	11,38
Trans	100	79,5	87,5	79,5	95	94,4	89,32	8,58
Tem	69,4	87,5	85,7	95,4	94,6	94,4	87,83	9,91
Total							**87,31**	**9,74**

Tabla I: Porcentajes de coincidencia de las palabras clave. ***Seleccionadas entre los 4 investigadores. Fuente propia.***

a. Criterio de control de la frecuencia de aparición de las palabras.
b. Criterio de control técnico: hay palabras menos aptas para ser localizadas por un sistema de reconocimiento automático (por ejemplo, monosílabos).
c. Criterio de control del vocabulario: hay palabras que necesariamente siempre irán seguidas de una preposición (ej.- "El Madrid" para referirse al equipo de fútbol).
d. Criterio de control semántico (Maniez, 1992): la semántica de una palabra suele ser múltiple, pero para nuestro trabajo debe quedar perfectamente acotada.

4. *Fichado de palabras clave en toda muestra*: se distribuyó la muestra total de noticias de forma que cada investigador analizara una parte de las noticias de cada tema. A partir de aquí, se generaba una "Ficha de vaciado de palabras" para cada noticia analizada. En esa ficha se identificaba numéricamente la noticia analizada, la cadena de emisión, locutores, campo temático, fecha, duración, el investigador que realizaba el análisis y, por supuesto, las palabras clave seleccionadas.
5. *Construcción de un Inventario-hipótesis*: el protocolo para entrar palabras clave en un primer inventario consistió en seleccionar de forma

acumulativa las tres palabras de mayor frecuencia de cada ficha, evitando repeticiones. En caso de que las palabras clave registradas en una ficha ya hubieran sido recogidas en el Inventario, se escogerían exclusivamente las no presentes. Este proceso continuó hasta que 5 fichas de vaciado se sucedían para un mismo tema sin que apareciese en ellas una sola palabra clave nueva para el inventario. El resultado de todo este trabajo fue la primera versión del inventario de palabras clave, al que denominamos a partir de ese momento "*Inventario-hipótesis"*, en tanto que cumplía perfectamente las condiciones de la hipótesis de trabajo: estaba fundamentado, tenía coherencia interna y podría ser contrastado utilizando procedimientos diversos.

4. *Verificación y depuración iterativa del inventario-hipótesis*: una vez preparada esta primera versión del inventario de palabras clave, era necesario ponerlo a prueba de algún modo que nos permitiese evaluar la capacidad clasificadora real del conjunto de palabras clave que habíamos escogido para cada una de las 15 categorías temáticas. Y estas pruebas nos debían permitir, además, desarrollar un proceso iterativo de mejora del inventario, depurándolo o ampliándolo en función de los resultados obtenidos. Este trabajo se realizó mediante el procedimiento que decidimos denominar como "*Test de eficacia"* y que consistía en utilizar el Inventario-hipótesis para un procedimiento que reproducía el funcionamiento del futuro sistema de reconocimiento sonoro; es decir, que buscaba en un número determinado de noticias todas las palabras clave del inventario presentes y evaluaba, luego, en función del campo temático al que pertenecían estas palabras clave, en qué campo temático quedaba clasificada cada noticia; y si este campo era, o no, el correcto.

   Una noticia quedaba clasificada en un campo temático en el momento en que la suma de apariciones de palabras clave de uno de los 15 campos predefinidos superaba la suma de todas las pertenecientes a cualquier otro campo temático. Se llevaron a cabo dos tests sucesivos de eficacia para verificar si las palabras seleccionadas clasificaban adecuadamente sendas muestras de noticias y, después de cada test, se hicieron las correspondientes correcciones

   Para desarrollar este test de eficacia se seleccionaron aleatoriamente de entre la muestra total del estudio, un paquete de 20 noticias que no habían sido todavía analizadas (ver Tabla II "Test de eficacia-1"):

Test de eficacia - 1: testeo de 20 noticias seleccionadas aleatoriamente de las no analizadas ínter-subjetiva o subjetivamente:

- Noticia 1 ("Societat"): NO SE CLASIFICA
  A) Número de palabras de "Trànsit": 4
  B) Número de palabras de "Societat": 4 ("policia" de 2º nivel con "Conflicte social")
  C) Número de palabras de "Economia i negocis": 2
- Noticia 2 ("Treball"): SÍ SE CLASIFICA
  A) Número de palabras de "Treball": 5
  B) Número de palabras de "Crònica internacional": 1
  C) Número de palabras de "Crònica política": 1
  D) Número de palabras de "Conflicte social": 1
- Noticia 3 …

(etc.)

Tabla II: Plantilla de resultados del Test de eficacia – 1. Fuente propia.

Este test también servía para hacer un primer diagnóstico de la integración estructural de las categorías temáticas en la noticia. Por ejemplo, el campo temático "Ciència i tecnologia" no contemplaba áreas como "arqueología" o "espacio exterior", o que las categorías "Conflicte Social" y "Societat" iban a necesitar palabras específicas que sólo sirvieran para diferenciarlas entre sí. Las propuestas concretas de modificación para el inventario-hipótesis pueden ser consultados en el informe del CAC. Sin embargo, siguiendo un criterio de máxima economía de palabras y basándonos en la teoría de los campos semánticos, finalmente sólo se añadieron al inventario las siguientes palabras clave: "CSIC", "descoberta", "nord-americà", "educador", "trasplantament", "programa", "graella", "pertorbació" y "bufar", y se decidió la necesidad de hacer otro test. Para desarrollar un segundo test de eficacia, que denominamos "Test de eficacia-2" se procedió del mismo modo que para el primer test, pero con una muestra bastante mayor y la presencia de todos los campos temáticos: muestra estratificada de 3 noticias por tema (45 en total). Las palabras que resultaron de este segundo test sí fueron incorporadas al Inventario en su totalidad (véase Rodríguez *et al*., 2006).

5. Depuración y corrección del inventario-hipótesis con procedimientos lexicométricos: los procedimientos de selección y testeo de palabras clave seguidos hasta ese momento nos habían permitido seleccionar conjuntos de palabras exclusivamente desde criterios semánticos. Esto nos garantizaba la diversidad de palabras clave para cubrir cualquiera de los campos semánticos contenidos en los 15 temas, pero no validaba el criterio *lexicométrico* que utiliza la máquina, y la posibilidad de incluir otras palabras bajo ese criterio únicamente (por asociación automática, por recursos estilísticos, por rutinas productivas, etc.). Para ello, realizamos un análisis *lexicométrico* de frecuencias comparadas orientado a localizar mediante procedimientos estadísticos nuevas palabras clave

no detectadas con el método anterior. Hablamos de una comparación sistemática y proporcional entre el número de apariciones de cada palabra dentro del conjunto de noticias de un campo temático, y el número de apariciones de esta misma palabra en todo el resto de campos temáticos de la muestra. Este nuevo proceso de análisis y mejora del inventario de palabras se hizo ya a partir de la transcripción de toda la muestra audiovisual (698 noticias) a texto escrito, lo que nos permitió hacer una gestión estadística de esta gran masa de palabras utilizando el software lexicométrico "Concordance". Según la ratio de aparición de cierta palabras, ésta sería incluida en el Inventario si:

1. Aparecía en el análisis un número de veces equivalente como mínimo al 10% de las noticias del campo temático que le correspondía y repartidas en más de una noticia
2. Si respondía a la siguiente ponderación:

$$E_c = \frac{(N - n_c) \cdot R_c}{n_c}$$

Siendo:
N: Número total de noticias en todos los campos temáticos
Rc: Número de repeticiones de la palabra clave "x" en las noticias de un campo temático concreto
Rr: Número total de repeticiones de la palabra clave "x" en las noticias de los campos temáticos restantes (excluyendo el campo analizado)
nc: Número de noticias del campo temático analizado

De modo que (**Ec**) fuese tres veces superior al número total de repeticiones de esta misma palabra clave en el resto de campos temáticos (**Rr**). Es decir si:

$$(E_c > 3 \cdot R_r)$$

El resultado final de esta tarea fueron las 15 listas de palabras que constituyeron el INVENTARIO-HIPÓTESIS y que fueron sometidas a contrastación estudiando su capacidad para clasificar noticias en los 15 campos temáticos predefinidos, utilizando procedimientos de análisis discriminante.

5. RESULTADOS Y DISCUSIÓN

Una vez cerrado el inventario hipotético, se preparó un fichero de datos con el número de apariciones y repeticiones de todas las palabras del inventario en la muestra completa

de noticias. Se pretendía probar estadísticamente la capacidad del Inventario-Hipótesis para contrastar la hipótesis:

Cada uno de los conjuntos de palabras clave actúa como grupo de variables con capacidad de discriminar un campo temático, estando estas asociadas a los 14 campos temáticos restantes.

Así, aplicando el análisis discriminante revisamos cada uno de los 15 conjuntos seleccionados y pudimos contrastar estadísticamente qué palabras tenían una capacidad discriminadora significativa para diferenciar el grupo palabras del campo temático al que pertenecían de los 15 restantes, qué otras no y en qué grado. A partir de este análisis fue posible definir, finalmente los 15 grupos de palabras clave que deberían ser utilizadas en el entrenamiento de un prototipo de reconocimiento automático.

Para realizar el análisis trabajamos con el paquete estadístico SPSS. Se contrastaba la presencia de los grupos de palabras temáticas del lexicón con el tema asignado a la noticia por parte del CAC. El estadístico utilizado fue la Lambda de Wilks, que contrasta las varianzas medias de las variables:

Varianza total = varianza "dentro de grupos" + varianza "entre grupos"
= proporción de variabilidad no explicada "entre grupos".

Por tanto, cuanto menor fuera el valor de la Lambda de Wilks (Λ), más grande sería el estadístico "F", mayor la varianza y, en consecuencia, más discriminadora la palabras clave en cuestión. Cada nueva variable se evaluó según su contribución a diferenciar el grupo teniendo en cuenta la actuación de la anterior o anteriores combinadas. Por eso la "Λ"$_s$ (el valor de Wilks antes de añadir una nueva variable) debía ser mayor que la "Λ_{s+1}" (el valor después de añadirla). De la relación entre ambas obteníamos el incremento de discriminación temática en el valor F (test de F). Y por tanto, era el valor "F" el que asignaba la ponderación de cada palabras clave, y el que dirigía nuestras decisiones sobre su bondad dentro del conjunto. Finalmente, las listas de palabras clave se configuraron según su mayor o menor incremento a la discriminación de cada grupo-hipótesis respecto del resto.

De un total de 604 palabras que tenía el Inventario-Hipótesis, se descartaron 180, quedando únicamente 424. En primer lugar, en la categoría *"Temps"* se seleccionaron 17 palabras de un total de 27, no porque hubiera diez que no funcionaban, sino porque con 17 el éxito del conjunto-hipótesis era muy alto y ya no mejoraba sensiblemente con la inclusión de más palabras. Naturalmente, se trata de una categoría bien definida y de léxico muy homogéneo. Lo mismo ocurrió en el caso de *"Tràns*i*t"* (11 palabras seleccionadas de un total de 20). Diferente es el caso de *"Treball"* (25 palabras de 28) y *"Sanitat"* (32 de 39), ya que a pesar de tener un léxico específico, sus campos semánticos pueden llegar a solaparse con los de *"Societat"* e incluso con *"Conflicte Social"*. Precisamente, sólo 29 palabras de un total de 65 fueron escogidas para la categoría *"Societat"*, que realmente configura un conjunto de palabras muy diverso y es la categoría que peor funciona.

La categoría *"Conflicte Social"* incluye el campo semántico -violencia social- y por tanto está mucho mejor definida (32 de 42 palabras fueron elegidas). Otras categorías que pueden confundirse son *"Crònica Política"* y *"Crònica Internacional"*, primero porque el criterio que define a los dos grupos es diferente (temático y geográfico, respectivamente) y segundo porque *"Crònica Internacional"* prioriza los aspectos políticos y de forma particular las actuaciones de la política nacional en ámbitos internacionales. Se escogieron 42 palabras clave de 65 para *"Crònica Internacional"*, la mayoría de las cuales son nombres de países y sus gentilicios, y 36 de 45 para *"Crònica Política"*.

Las categorías de *"Esports"* (51 palabras seleccionadas de 74 iniciales), *"Ciència y tecnologia"* (33 seleccionadas de 48), *"Mitjans de Comunicació"* (20 de 27, *"Educació i ensenyament"* (20/32) y *"Art i Cultura"*, son categorías con diversos campos semánticos perfectamente delimitados, por lo que consideramos que en futuros trabajos sus palabras clave deberían subdividirse en grupos para cada uno de esos campos y realizar un estudio suplementario. También *"Medi Ambient"* tiene muchos campos semánticos, demasiados para ser contemplados en palabras por lo que, a pesar de ser una categoría bien definida, observamos que muchas de sus palabras clave se solapaban entre sí e incluso con palabras de la categoría *"Temps"*. Por último, observamos que la categoría *"Economia i negocis"* estaba perfectamente definida y con un grupo de 45 palabras de las 62 propuestas se conseguía un funcionamiento muy eficiente. Un rasgo a resaltar es que muchas palabras clave se convirtieron en tales por su significado figurado (ej.- "bolsa" es una palabra cuyo significado figurado en el mercado bursátil resulta clave para la categoría).

A partir de aquí, sólo nos quedaba hacer el desglose de las palabras según sus variaciones fónicas para poder entrenar al sistema de reconocimiento de forma precisa. El inventario definitivo resultante fue de unas 1000 palabras clave distribuidas en los 15 campos temáticos investigados del siguiente modo:

1. Temps: *Anticicló, Boira, Bufar, Calor, Meteorològica, Nevada, Nevar, Nuvolada, Núvols, Pertorbació, Ploure, Pluja, Precipitació, Ruixat, Temperatura, Tempesta, Tronada.*
2. Trànsit: *AP-7, Autopista, Carretera, Circulació, Circular, Conductor, Cotxe, Operació.sortida, Peatge, Trànsit, Vehicles.*
3. Treball: *Atur, Augment, CGT, Comissions.obreres, Construcció, Contractar, Contracte, Desocupats, Empresa, Fàbrica, Industrial, Laboral, Negociació, Obrer, Parats, Recursos.humans, Salari, Salarial, SEPI, Sindical, Sindicat, Treball, Treballador, Vaga, Xifres.*
4. Esports: *ACB, Amistós, Atlètic, Barça, Bàsquet, Betis, Blanc-i-blau, Blaugrana, Campió, Campionat, Champions, Copa, Cursa, Davanter, Derbi, Derrota, Disputar, Divisió, Entrenador, Entrenament, Entrenar, Esport, Federació, Fitxatge, Futbol, Golejador, Gran.premi, Guanyar, Jugadors, Jugar, Lesió, madrid Al/Del/El, Marcador, Montmeló, Natació, NBA, Olímpic.*
5. Sanitat: *Afectat, Anticonceptiu, Assistència, Càncer, Cirurgia, Clínic, De.sang, Donació, Donant, Droga, Embaràs, Epidèmia, Hospital, Hospitalitzar, Infermera, Legionel·la, Legionel·losi, Malalt, Malaltia, Medicina, Metge, Operacions, Pacient, Quiròfan, Salut, Sanitari, Sanitat, Símptoma, Tabac, Teràpia, Transfusió, Urgència.*
6. Ciència i tecnologia: *ADN, Agència.espacial, Aparell, Astronauta, Ciència i tecnologia, Científic, Clonació, Coet, CSIC, Descobriment, Descobrir, Digital, Electrònic, Embrió, Estudi, Genètic, Geològic, Innovador, Investigació, Investigador, La.terra, Laboratori, Missió, NASA, Òrbita, Ordinador, Recerca, Robot, Satèl·lit, Sensor, Sonda, Tecnologia, Teoria, Troballa.*
7. Mitjans de Comunicació: *Analògic, Audiència, Comunicació, Emissora, Internet, Nova.tecnologia, Oient, Pantalla, Programa, Ràdio, Senyal, Societat.de.la.informació, TDT, Telecomunicació, Telefilm, Telespectador, Televisió, Xarxa.*
8. Art i cultura: *Actor, Adaptació, Artista, Bolliwood, Cantant, Castellà, Cel·luloide, Cine, Cinema, Col·lecció, Còmic, Curtmetratge, Dansa, Escriptor, Espectacle, Estrella, Estrenar, Exhibició, Exposició, Festival, Fòrum, Hollywood, Homenatge, Literatura, Llengua, Llibre, Música, Musical, Novel·la, Obra, Orquestra, Pel·lícula, Poesia, Protagonista, Recinte, Sant.joan[1], Teatre.*
9. Medi Ambient: *Aigües, Allau, Ambiental, Boscos, Calorós, Contaminació, Cremar, Dotació, Ecologista, Epicentre, Flames, Forestal, Hectàrea, Huracà, Inundació, Medi.ambient, Natura, Nuclear, Onada de fred, Onada de calor, Platja, Port de muntanya, Provocar, Residu, Temporal, Terratrèmol, Terreny, Tifó, Tòxic, Vegetació, Vessament.*
10. Economia i negocis: *Ibex, Agricultura, Alcista, Ave, Baixat, Banc.Santander, Barril, Borsa, Comerç, Consumidor, Crèdit, Dèficit, Dòlar, Dow.jones, Economia i negocis, Econòmic, Encarir, Estalviar, Euro, Esportar, Finançament, Financer, Gasolina, Hoteler, Indústria, Inflació, Infraestructures, Ingressos, Invertir, Mercat, Multinacional, Nasdaq, Negoci, Pagament, Pagesos, Per.cent, Petroli, preu.El/Els/De/Del/Dels, Rebaixa, Regadiu, Sector, TGV, Turisme, Turista, Turístic.*
11. Crònica Internacional: *Al.kaida, Ambaixada, Atac, Bagdad, Bolívia, Boston, Branderburg, Cuba, Demòcrata, Estats.de.la.unió, Estats.units, Europa, Guerra, Hostatge, Indígena, Crònica Internacional, Iraq, Islamista, Israel, Londres, Marroc, Mèxic, Moscou, Najab(a), Nord.amèrica, Palestina, Paris, Partit, President.rus, Primer.ministre, Regne.Unit, República, Revolta, Russia, Segrest, Soldats, Terrorista, Tokio, Txetxènia, Uzbekistan, Xiïta.*
12. Crònica Política: *Audiència.nacional, Batasuna, Cimera, Comissió, Congrés, Consell, Conseller, Constitució, Convergència, Debat, Desactivar, Diputat, Elecció, Electoral, Esquerra, Estatut, ETA, Etarra, Euro.regió, Executiu, Explosiu, Forces.armades, Govern, Ministre, Nacionalista, Oposició, Parlament, Polític, PP, President, PSC, Referèndum, Republicà, Socialista, SOE.*
13. Societat: *Accident, Aparcar, Autòpsia, Avaria, Carrer, Cristià, Destrossa, Domèstic, Dona, Guàrdia, Habitatge, Immigrant, Implicar, Judici, Jutge, Legal, Marit, Mobilitat, Multa, Pisos, Presumpte, Religió, Sentència, Succés, Temerari, Vandalisme, Xocar.*
14. Conflicte Social: *AENA, Agents, Aldarull, Barri, Calabós, Col·lectiu, Coordinadora, Demanda, Denunciar, Eskin, Feixista, Mobilització, Municipal, Okupa, Pastera, Policial, Prostituta, Protestar, Queixa, Reclamar, Regularització, Reivindicar, Sense papers, Signatures, Soroll, Sortir al carrer, Subsaharià, Suïcidar, Veïna.*
15. Educació i Ensenyament: *Alumne, Assignatura, Bressol, De/Dels/Entre.zero.a/als /i tres, Docent, Educació i ensenyament, Educador, Educatiu, Ensenyament, Escola, Escolar, ESO, Estudiant, Examen, Guarderia, Infantil, Mestre, Pedagògic, Professor, Selectivitat.*

Tabla III. Inventario de palabras clave temáticas. Fuente propia.

En la siguiente tabla se pueden ver los resultados obtenidos en el análisis estadístico (Tabla IV):

	CLASIFICA	EXPLICA	DISTANCIA
CIÈNCIA I TECNOL.	82,5	74,3	7,311
MITJANS DE COM.	64,1	54,7	4,789
SANITAT	87,1	72,3	7,818
TEMPS	87,9	85,8	11,555
MEDI AMBIENT	60,5	54,8	4,575
TRANSIT	70,4	64,0	6,91
CRÒNICA POLÍTICA	57,3	51,0	2,871
ECONOMIA I NEG.	69,9	61,6	5,137
CRÒNICA INT.	67,4	62,4	5,184
ESPORTS	91,7	89,3	13,006
SOCIETAT	45,7	31,1	1,967
CONFLICTE SOCIAL	67,5	57,1	5,046
TREBALL	88,2	72,4	7,526
EDUCACIÓ I ENS.	69,0	66,6	7,069
ART I CULTURA	69,6	61,6	4,646
Mitjanes:	**71,92**	**63,92**	**6,36**

Tabla IV. Resultados globales sobre el análisis discriminante del inventario-hipótesis. Fuente propia.

En ella relacionamos: a) los resultados previstos de porcentaje de clasificación de noticias que permite cada conjunto de palabras seleccionadas, b) el porcentaje de varianza que explica cada función discriminante y c) la distancia entre centroides de los pares de grupos analizados (suma de las funciones de centroides de cada par de grupos), pudiendo observar lo siguiente:

1. El porcentaje medio de palabras que el procedimiento de análisis aplicado prevé que podrían ser clasificadas adecuadamente a partir de este listado de palabras clave supera el 70 %.
2. El porcentaje medio del fenómeno explicado por el conjunto de los listados de palabras clave es cercano al 65 %.
3. La distancia media entre centroides está por encima de 6 puntos en un fenómeno en el cual el intervalo de referencia mínimo para conseguir diferenciar es la aparición (1) o la no aparición (0) de una palabra en una determinada noticia.

Podemos afirmar, en consecuencia, que los resultados obtenidos resultan globalmente muy aceptables desde el punto de vista estadístico. El campo temático más problemático resultó ser el de *"Societat"*. Tal y como se puede observar en la Tabla IV su diferencia respecto al resto de grupos analizados es notable. La eficiencia del conjunto de palabras seleccionadas como instrumento de clasificación de este campo está casi un 30 % por debajo de los otros campos. Este resultado bajo aislado apunta a la indefinición de la categoría temática.

El problema del reconocimiento automático de palabras en habla continua no es una cuestión menor; es posible que las necesidades para la puesta en marcha de este sistema influyan de forma no esperada en las características del propio listado de palabras y sea necesario revisarlo en una cierta medida. De hecho, mientras no dispongamos de las primeras experiencias y resultados con un *software* de reconocimiento vocal automático entrenado para reconocer este inventario no estaremos en condiciones de proponer un algoritmo fiable de clasificación.

Otro de los problemas que con seguridad aparecerá en la próxima fase de desarrollo de este programa de investigación es la "suciedad sonora" que proviene de las locuciones audiovisuales (cambios de locutor en la misma unidad informativa, músicas de fondo, ambientes sonoros, locuciones defectuosas, locuciones superpuestas, etc.). Estos factores probablemente obligarán a buscar nuevas estrategias para que los inventarios de palabras clave incorporen criterios de robustez acústica adaptados a las necesidades específicas del sistema de reconocimiento.

Finalmente, prevemos también la aparición de dificultades vinculadas a la problemática de la proximidad acústica entre determinadas palabras clave cuando esta esté asociada a la variabilidad natural de las locuciones.

6. CONCLUSIONES

Para esta investigación hemos desarrollado un protocolo específico de análisis y hemos estudiado una muestra de 698 noticias televisivas que contenían 23.298 palabras diferenciadas y configuraban una masa total de más de 160.000 vocablos si incluimos las repeticiones. Y se ha realizado, también, una relectura de la teoría del discurso, aplicándola al problema de la clasificación temática de las noticias, así como una revisión del conocimiento disponible en torno a los métodos y sistemas de catalogación y análisis de contenidos automáticos y no automáticos.

Como resultado de toda esta tarea, los resultados son el propio inventario de 1.000 palabras clave que hemos presentado y que configura el nuevo conocimiento generado. Este listado nos permitió pasar a la siguiente fase de nuestro proyecto sobre la preparación de un sistema automático para la clasificación temática de noticias, que es la localización de rasgos sonoros que faciliten la segmentación automática de las noticias (Mas Manchón, 2011, Tesis Doctoral). No obstante, creemos que en este trabajo, por encima de un lexicón que debe ser revisado periódicamente, se ha validado una metodología para la localización de palabras clave temáticas.

La investigación desarrollada en este estudio sobre instrumentos automáticos de reconocimiento, clasificación e indización y la tarea desarrollada de selección de palabras clave para clasificar noticias a partir de unos criterios semántico-funcionales nos ha llevado a la conclusión de que es perfectamente viable orientar esta línea de trabajo hacia la automatización del análisis de contenido de la información audiovisual. Así,

entendemos que la implementación de un sistema automático de análisis de contenido puede basarse en procedimientos de reconocimiento automático de palabras clave (*word spottig*). La base de este procedimiento consiste en una orientación específica de la construcción de los inventarios de palabras clave hacia este objetivo específico.

NOTAS

[1] El Consell de l'Audiovisual de Catalunya (CAC) se constituyó en 1997 como organismo asesor del gobierno de la Generalitat de Catalunya con el objetivo de garantizar el cumplimiento de las normas sobre los contenidos en televisión y radio.

[2] El Laboratorio de Análisis Instrumental de la Comunicación (LAICOM) es un centro de investigación del Dpto. de Comunicación Audiovisual y Publicidad II de la Universidad de Barcelona.

[3] Efectivamente, no todas las investigaciones desde el campo de la ingeniería se han limitado al criterio lexicométrico para la elección de palabras clave, como se demuestra en Dimitrova *et al*. (1999) y Yang *et al*. (2004), pero siempre son criterios suplementarios al análisis *lexicométrico*. En todo caso, no hemos encontrado abordajes que hagan un análisis de la función discursiva de la palabra.

8. BIBLIOGRAFÍA

ABBERLEY, D.; KIRBY, D.; RENALS, S. y ROBINSON, T. The This broadcast news retrieval System. Sheffield: University of Sheffield. Department of Computer Science, 2006, [en línea], URL: <http://svr-www.eng.cam.ac.uk/~ajr/esca99/Abberley.pdf> [Consulta: 17 de marzo de 2006].

ARIKI, Y. y MATSUURA, K. Automatic Classification of TV News Articles based on telop Character Recognition**.** Japón: *Ryukoku University, Department of Science and Technology, 1999,* [en línea], URL: <http://ieeexplore.ieee.org/iel5/6322/16898/00778210.pdf?isnumber=&arnumber=778210> [Consulta: 8 de marzo de 2011].

BERTALANFFY, L.V. T*eoría General de Sistemas.* México: Fondo de Cultura Económica, 1968.

BROWN, G. y YULE, G. *Análisis del Discurso*. Madrid: Visor, 1993.

CEBRIÁN HERREROS, M. *Información radiofónica. Mediación técnica, tratamiento y programación.* Madrid: Síntesis, 1994.

CHARAUDEAU, P. *El discurso de la información: la construcción del espejo social*. Barcelona: Gedisa, 2003.

DIMITROVA, N. Multimedia Content Analysis and Indexing for Filtering and Retrieval Applications. En Informing Science. Special issue on Multimedia Informing Technologies, Part 1. Vol. 2, nº 4, 1999, [en línea] URL: <http://inform.un/Articles/Vol2/v2n4p87-100.pdf> [Consulta: 20 de mayo de 2005].

DONOUGH, J.; SIU, M. y GISH, H. Reducing word error rate on conversational speech from the Switchboard corpus. En ICASSP 95, 53(56), 1995, [en línea]. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00479271> [Consulta: 8 de marzo de 2011].

LIU, Z.; HUANG, J. y WANG, Y. Classification of TV programs based on audio information using Hidden Markov Model. Nueva York: Department of Electrical Engineering. Polytechnic University, Brooklyn, 1998a, [en línea]. URL:

<http://ieeexplore.ieee.org/iel4/5958/15944/00738908.pdf?arnumber=738908> [Consulta: 10 de octubre de 2006].

LIU, Z.; WANG, Y. y CHEN, T. Audio Feature Extraction and Analysis for Scene Segmentation and Classification. Nueva York: Polytechnic University, Brooklyn. Carnegie Mellon University, Pittsburgh, 1998b, [en línea]. URL: <http://portal.acm.org/citation.cfm?id=302295> [Consulta: 8 de marzo de 2011].

MANIEZ, J. *Los lenguajes documentales y de clasificación. Concepción, construcción y utilización en los sistemas documentales.* Madrid: Pirámide, 1993.

MARTÍNEZ ALBERTOS, J.L. *Curso General de redacción periodística*. Madrid: Paraninfo, 1998.

MAS MANCHÓN, LL. *Modelos Entonativos para la Segmentación Automática de los Programas Informativos en Unidades-Noticia*. Tesis Doctoral. Barcelona: Dpto. Comunicación Audiovisual y Publicidad II, UAB, 2011.

MILLER, M.M. y RIECHERT, B.P. 1994, [en línea] URL: <http://excellent.com.utk.edu/~mmmiller/pestmaps.txt> [Consulta: 11 de mayo de 2011].

MIZOGUCHI, R.; TSUNEKAWA, T. y YAMASHITA, Y. Topic Recognition for News Speech based on Keyword Spotting. I.S.I.R. Osaka University. 1-8 Mihogaoka, Ibaraki-shi. Osaka, 567-0047, Japón, Ritsumeikan University. 1-1-1 Noji- Higashi. Kusatsu-shi. Shiga, 525-8577, Japón: 5th International Conference on Spoken Language Processing (ICSLP '98), Sydney, 3, 1998, [en línea] URL: <http://www.slp.is.ritsumei.ac.jp/~yama/pubs/icslp98.pdf> [Consulta: 8 de marzo de 2011].

NAKAMURA, Y. y KANADE, T. Semantic Analysis for Video Contents Extraction-Spotting by Association in News Video. ACM Multimedia – Electronics Proceedings. Crowne Plaza Hotel, Seattle, USA, 8-14 noviembre, 1997, [en línea] URL: <http://www.image.esys.tsukuba.ac.jp/~yuichi/online-paper/ACM1997/main.html> [Consulta: 27 de mayo de 2005].

NAPHADE, M.R. y HUANG, T.S. Semantic filtering of Video Content. 2005. [en línea]. URL: <http://www-scf.usc.edu/~csci586/papers/video/BPtemp13402.pdf> [Consulta: 18 de mayo de 2005].

NAPHADE, M.R.; KOZINTSER, I.V. y HUANG, T.S. A factor Graph Framework for Semantinc Video Indexing**,** 2004**.** [en línea] URL: <http://www.kozintsev.net/papers/jounal_02.pdf> [Consulta: 8 de marzo de 2011].

PINTO, M. *Manual de Clasificación Documental*. Madrid: Síntesis, 1997.

RENALS, S.; ABBERLEY, D.; KIRBY, D. y ROBINSON, T. Indexing and Retrieval of Broadcast News**.** IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing, 13-15 de septiembre, 1999, [en línea] URL: <http://homepages.inf.ed.ac.uk/srenals/pubs/1999/mmsp99-54/mmsp99-54.html> [Consulta: 8 de marzo de 2011].

RODRÍGUEZ BRAVO, A. *et al.* Clasificador Automático de Información Sonora. Proyecto financiado por el CAC. Barcelona: Depósito del Laicom, Edifici I, UAB, Bellaterra, Barcelona, y en el Consell de l'Audiovisual de Catalunya, 2006.

RODRÍGUEZ BRAVO, A. La investigación aplicada: una nueva perspectiva para los estudios de recepción*. Quaderns de Comunicación i cultura*, 2003, nº 30, p. 17-36.

RODRÍGUEZ BRAVO, A. Fundamentos para una teoría de la eficacia comunicativa. Actas del Congresso Brasileiro de Ciências da Comunicação (Intercom). Natal, RN. Del 2 al 6 de septiembre, 2008, [en línea]. URL: <http://www.intercom.org.br/papers/nacionais/2008/resumos/R3-0572-1.pdf> [Consulta: 8 de marzo de 2011].

SHANON, C.E. y WEAVER, W. *Teoría Matemática de la Información*. Madrid: Forka, 1981.

SIERRA BRAVO, R. *Técnicas de investigación social.* Madrid: Paraninfo, 2001.

VAN DIJK, T.A. *La noticia como discurso: comprensión, estructura y producción de la información*. Barcelona: Paidós, 1990.

WIENER, N. *Cibernética y sociedad*. Buenos Aires: Sudamericana, 1969.

YANG, C.; DONG, M. y FOTOHUI, F. Learning the semantics in image retrieval – A natural language processing approach. Estados Unidos: Computer Science Department, Wayne State University, 2004, [en línea]. URL: <http://ieeexplore.ieee.org/iel5/9515/30163/01384934.pdf?arnumber=1384934> [Consulta: 8 de marzo de 2011].