MEMÒRIES DE LA SOCIETAT CATALANA DE TERMINOLOGIA, 1

Jaume Martí i Marina Salse (coord.)

LA TERMINOLOGIA I LA DOCUMENTACIÓ: RELACIONS I SINERGIES

UNIVERSITAT DE BARCELONA
Facultat de Biblioteconomia i Documentació

scatterm
SOCIETAT CATALANA DE TERMINOLOGIA
Filial de l'Institut d'Estudis Catalans

INSTITUT D'ESTUDIS CATALANS MCMVII

La terminologia i la documentació: relacions i sinergies

Mesa inaugural de la VII Jornada de la SCATERM.
D'esquerra a dreta, Jaume Martí, president de la SCATERM; Salvador Alegret, vicepresident de l'Institut d'Estudis Catalans; Gemma Fonrodona, vicerectora d'Estudiants i Política Lingüística de la Universitat de Barcelona, i Cristóbal Urbano, degà de la Facultat de Biblioteconomia i Documentació de la mateixa universitat

Vista dels assistents a la VII Jornada de la SCATERM, tinguda el dia 29 de maig de 2009 a l'Aula Jordi Rubió i Balaguer de la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona

SOCIETAT CATALANA DE TERMINOLOGIA
FILIAL DE L'INSTITUT D'ESTUDIS CATALANS

MEMÒRIES DE LA SOCIETAT CATALANA DE TERMINOLOGIA, 1

JAUME MARTÍ I MARINA SALSE
(coord.)

La terminologia i la documentació: relacions i sinergies

Actes de la VII Jornada de la SCATERM:
«Terminologia i documentació»
(Facultat de Biblioteconomia i Documentació,
Universitat de Barcelona, 29 de maig de 2009)

BARCELONA, 2010

Biblioteca de Catalunya. Dades CIP

Jornada de la SCATERM (7a : 2009 : Barcelona)
La Terminologia i la documentació: relacions i sinergies : actes de la VII Jornada de la SCATERM "Terminologia i documentació". — (Memòries de la Societat Catalana de Terminologia ; 1)
Jornada celebrada a la Facultat de Biblioteconomia i Documentació, Universitat de Barcelona, 29 de maig 2009. — Bibliografia
ISBN 9788492583867
I. Martí, Jaume (Martí Llobet), ed. II. Salse, Marina, ed. III. Societat Catalana de Terminologia V. Universitat de Barcelona. Facultat de Biblioteconomia i Documentació V. Títol
VI. Col·lecció: Memòries de la Societat Catalana de Terminologia ; 1
1. Documentació — Congressos 2. Terminologia — Congressos
801.3:002(061.3)

Primera edició: febrer de 2010
Tiratge: 600 exemplars

Text revisat lingüísticament pel Servei de Correcció Lingüística de l'IEC

Compost per Anglofort, SA
Imprès a Limpergraf, SL

ISBN: 978-84-92583-86-7
Dipòsit Legal: B. 7350-2010

Taula

Organització 7

Participants 9

Programa de la VII Jornada de la SCATERM 11

Sigles emprades pels autors 13

Presentació,
per Jaume Martí 15

Sessió I
Ponència

Encontrar documentos a través de las palabras y de los enlaces,
per José L. Alonso Berrocal 19

Comunicacions

Invitació a l'estudi estadístic del llenguatge,
per Rogelio Nazar 47

Ús d'estratègies estadístiques per a l'extracció automàtica d'unitats terminològiques,
per Mercè Vázquez i Antoni Oliver 75

La documentació aplicada a la traducció jurídica,
per Eivor Jordà 85

El vocabulari de preservació i conservació del patrimoni documental,
per Maria Elvira 93

SESSIÓ II
Ponència

El futur de la informació acadèmica: Web semàntic / Web social, o tots dos?
per Lluís Codina 105

Comunicacions

Vocabulària: un multicercador temàtic,
per Xavier Albons, Pep Cara, Àngels Egea i Montserrat Lleopart 119

Terminologia i documentació 2.0,
per Jordi Chumillas, Ruth S. Contreras i Ricard Giramé 125

Balanç i conclusions de la VII Jornada de la SCATERM
per Marina Salse i Jaume Martí 135

Assistents a la VII Jornada 139

Organització

Institucions organitzadores:

Societat Catalana de Terminologia (SCATERM)
Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona (UB)

Amb la col·laboració de l'Institut d'Estudis Catalans, el Vicerectorat d'Estudiants i Política Lingüística (UB), el Vicerectorat de Política Científica (UB) i la Comissió de Dinamització Lingüística de la Facultat de Bibliotecomia i Documentació (UB)

Coordinadors:

Jaume Martí i Llobet
President de la Societat Catalana de Terminologia

Marina Salse Rovira
Facultat de Biblioteconomia i Documentació (UB)

Comitè organitzador:

Àngels Egea Puigventós
Societat Catalana de Terminologia

Núria Jornet Benito
Facultat de Biblioteconomia i Documentació (UB)

Josep M. Mestres i Serra
Societat Catalana de Terminologia

Marina Salse Rovira
Facultat de Biblioteconomia i Documentació (UB)

Participants

Xavier Albons Gomila
Serveis Lingüístics
Universitat de Barcelona

Salvador Alegret i Sanromà
Vicepresidència
Institut d'Estudis Catalans
Barcelona

José Luis Alfonso Berrocal
Departament d'Informàtica i Automàtica
Universitat de Salamanca

Miquel Centelles Velilla
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Jordi Chumillas i Coromina
Universitat de Vic

Lluís Codina Bonilla
Secció de Ciències de la Documentació
Universitat Pompeu Fabra
Barcelona

Àngels Egea i Puigventós
Serveis Lingüístics
Universitat de Barcelona

Maria Elvira i Silleras
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Gemma Fonrodona Baldajos
Vicerectorat d'Estudiants i Política Lingüística
Universitat de Barcelona

Ricard Giramé Parareda
Universitat de Vic

Eivor Jordà Mathiasen
Centre universitari ESTEMA
València

Rogelio Nazar
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

Jaume Martí i Llobet
Societat Catalana de Terminologia
Barcelona

Antoni Oliver González
Universitat Oberta de Catalunya
Barcelona

Marina Salse Rovira
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Cristóbal Urbano Salido
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Mercè Vázquez i Garcia
Universitat Oberta de Catalunya
Barcelona

VII Jornada de la SCATERM: «Terminologia i documentació»

Facultat de Biblioteconomia i Documentació,
Universitat de Barcelona, 29 de maig de 2009

Programa

9.30 h Inscripció de participants i lliurament de documentació

10.00 h *Inauguració de la Jornada*

Gemma Fonrodona Baldajos
Vicerectora d'Estudiants i Política Lingüística. Universitat de Barcelona

Salvador Alegret i Sanromà
Vicepresident de l'Institut d'Estudis Catalans

Cristóbal Urbano Salido
Degà de la Facultat de Biblioteconomia i Documentació. Universitat de Barcelona

Jaume Martí i Llobet
President de la Societat Catalana de Terminologia, filial de l'Institut d'Estudis Catalans

10.15 h *Ponència*
Encontrar documentos a través de las palabras y de los enlaces
José Luis Alonso Berrocal
Professor titular del Departament d'Informàtica i Automàtica. Universitat de Salamanca

11.30 h Pausa (cafè)

12.00 h *Comunicacions*
Invitació a l'estudi estadístic del llenguatge
Rogelio Nazar
Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra (Barcelona)

Ús d'estratègies estadístiques per a la recuperació automàtica d'unitats terminològiques
ANTONI OLIVER i MERCÈ VÁZQUEZ
Universitat Oberta de Catalunya (Barcelona)

La documentació aplicada a la traducció jurídica
EIVOR JORDÀ MATHIASEN
Centre universitari ESTEMA (València)

Vocabulari de preservació i conservació del patrimoni documental
MARIA ELVIRA
Universitat de Barcelona

13.20 h *Debat del matí*
Moderador: MIQUEL CENTELLES

15.30 h *Ponència*
El futur de la informació acadèmica: Web semàntic / Web social, o tots dos?
LLUÍS CODINA
Professor titular de la Secció de Ciències de la Documentació
Universitat Pompeu Fabra

17.00 h Pausa (cafè)

17.30 h *Comunicacions*
Vocabulària: un multicercador temàtic
XAVIER ALBONS, PEP CARA, ÀNGELS EGEA i MONTSERRAT LLEOPART
Serveis Lingüístics de la Universitat de Barcelona

Terminologia i Documentació 2.0
JORDI CHUMILLAS, RICARD GIRAMÉ i RUTH CONTRERAS
Universitat de Vic

18.15 h *Debat de la tarda*
Moderador: JAUME MARTÍ

18.45 h *Balanç i conclusions de la VII Jornada*
MARINA SALSE i JAUME MARTÍ

Sigles emprades pels autors

ACM	Association for Computing Machinery
AESLA	Asociación Española de Lingüística Aplicada
CPNL	Consorci per a la Normalització Lingüística
CSV	*comma-separated values* ('valors separats per comes')
DEM	*Diccionari enciclopèdic de medicina* (Enciclopèdia Catalana)
DIEC	*Diccionari de la llengua catalana* (Institut d'Estudis Catalans)
DRAE	*Diccionario de la lengua española* (Real Academia Española)
EMNLP	*empirical methods in natural language processing* ('mètodes empírics en el processament de llengües naturals')
FLINS	*fuzzy logic and intelligent technologies in nuclear science* ('lògica difusa i tecnologies intel·ligents en la ciència nuclear')
GDLC	*Gran diccionari de la llengua catalana* (Enciclopèdia Catalana)
HITS	*hypertext induced topic selection* ('selecció de temes a partir de l'hipertext')
HTML	*hypertext markup language* ('llenguatge d'etiquetatge d'hipertext')
HTTP	*hypertext transfer protocol* ('protocol de transferència d'hipertext')
IDF	*invers document frequency* ('freqüència inversa de document')
IEC	Institut d'Estudis Catalans
IMS	Institut für Maschinelle Sprachverarbeitung
IRI	*internationalized resource identifier* ('identificador internacionalitzat de recursos')
IULA	Institut Universitari de Lingüística Aplicada
IULACT	corpus textual de l'Institut Universitari de Lingüística Aplicada
MP3	*Moving Picture Experts Group-1 audio layer-3* ('capa d'àudio-3 del Grup d'Experts en Imatges en Moviment-1')
MIT	Massachusetts Institute of Technology
OPAC	*online public access catalog* ('catàleg en línia d'accés públic')
OWL	*ontology web language* ('llenguatge web d'ontologies')
PC	*personal computer* ('ordinador personal')
PDA	*personal digital assistant* ('organitzador personal digital')

PDF	*portable document format* (‘format de document portàtil’)
PIM	*personal information manager* (‘gestor d’informació personal’)
PLN	*procesamiento del lenguaje natural* (‘processament del llenguage natural’)
RDF	*resource description framework* (‘marc de descripció de recursos’)
RDFS	*resource description framework schema* (‘esquema de marc de descripció de recursos’)
RI	*recuperación de información* (‘recuperació d’informació’)
RIF	*rule interchange format* (‘format d’intercanvi de regles’)
SCATERM	Societat Catalana de Terminologia
SEO	*search engine optimization* (‘optimització dels motors de cerca’)
SEPLN	Sociedad Española para el Procesamiento del Lenguaje Natural
SIGDAT	*special interest group on linguistic data and corpus-based approaches to natural language processing* (‘grup d’interès especial en dades lingüístiques i en l’aproximació al processament del llenguatge natural basat en corpus’)
SIGIR	*special interest group on information retrieval* (‘grup d’interès especial en la recuperació d’informació’)
SPARQL	*simple protocol and RDF query language* (‘llenguatge d’interrogació de protocol simple i marc de descripció de recursos’)
TF	*term frequency* (‘freqüència de terme’)
TVE	Televisión Española
UB	Universitat de Barcelona
UPC	Universitat Politècnica de Catalunya
UPF	Universitat Pompeu Fabra
URI	*uniform resource identifier* (‘identificador uniforme de recursos’)
URL	Universitat Ramon Llull
UVic	Universitat de Vic
VLC	*very large corpora* (‘corpus molt llargs’)
W3C	World Wide Web Consortium
WWW	*World Wide Web* (‘Web’)
XML	*extensible markup language* (‘llenguatge extensible de marcatge’)
XSL	*xtensible stylesheet language* (‘llenguatge de fulls d’estil extensible’)

Presentació

Jaume Martí
President de la SCATERM

El panorama actual del coneixement vinculat a la recerca, el configuren una gran varietat de disciplines i subdisciplines, en diguem *ciències*, *tècniques* o *tecnologies*, lligades als seus corresponents camps d'activitat professional. Aquest lligam prové en general del fet que al voltant dels coneixements de cada disciplina s'ha generat activitat professional; però no és rar el fet invers, de disciplines sorgides de les necessitats o conveniències d'una parcel·la professional. Tot plegat és l'estructura cognitiva i social de què ens hem dotat per abordar el coneixement i fer-lo socialment rendible.

El sentit de l'evolució històrica ha estat el de la separació o parcel·lació progressiva en aquests espais, a mesura que l'aprofundiment en els coneixements i la consegüent especialització ho requerien. Però també és cert que els contactes i els punts comuns entre diferents àmbits especialitzats han estat cada vegada més evidents i indefugibles. Podríem dir que els transvasaments de teories i de coneixements entre disciplines són part inherent del procés mateix i són en l'origen del fenomen de la transversalitat.

Segurament és en els vessants pràctics i aplicats allà on moltes disciplines aparentment allunyades es toquen, on les confluències o trobades són més necessàries i la col·laboració entre els professionals, més útils i fins i tot imprescindibles.

En aquest sentit, el cas de la terminologia i la documentació, dues disciplines amb un vessant aplicat molt important, és un dels més clars.

La terminologia estudia i explica des de la lingüística aquests elements lèxics especials que són els termes, i en el seu vessant aplicat els detecta en els textos, en fa reculls, hi introdueix propostes, etc. La documentació fa ús dels termes i els adapta per als seus fins de construcció de classificacions i recuperació de la informació dels textos i dels corpus textuals.

Des de la premissa de l'existència d'aquests vincles i de la unitat indissociable

que formen la documentació i la terminologia, mostrada ja en una primera jornada organitzada per l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra l'any 2000, la Jornada titulada «Terminologia i documentació», les actes de la qual apleguem ací, fou el marc de trobada per a desenvolupar el coneixement sobre els punts de contacte damunt esmentats mitjançant les aportacions dels estudiosos i els professionals que s'hi aplegaren al llarg de tot el dia 29 de maig de 2009 a la Universitat de Barcelona.

Amb una assistència d'una seixantena de professionals i estudiosos de la terminologia i la documentació, dues ponències, impartides per José Luis Alonso Berrocal i Lluís Codina, respectivament, van servir de pal de paller de les sis comunicacions que es descabdellaren durant el matí i la tarda que durà la VII Jornada.

Sessió I

Ponència

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 19-43 DOI: 10.2436/15.2503.02.1

Encontrar documentos a través de las palabras y de los enlaces

José L. Alonso Berrocal
Departamento de Informática y Automática
Universidad de Salamanca

Resumen

Esta ponencia se centra en la recuperación de palabras y enlaces para encontrar documentos. En ella se exponen los métodos de recuperación de información, tanto los métodos teóricos como la indización práctica, y se resumen sus resultados. También se explica ampliamente en qué consiste el modelo vectorial de recuperación de información y, finalmente, se habla de las técnicas de recuperación en la Web y su relación con el llamado *spamdexing*, que es la capacidad de conseguir ocupar las primeras posiciones de los motores de búsqueda.

Palabras clave: *spamdexing*, indización, modelo vectorial de recuperación de información, recuperación de información.

Resum: ***Trobar documents per mitjà de les paraules i dels enllaços***

Aquesta ponència se centra en la recuperació de paraules i enllaços per a trobar documents. S'hi exposen els mètodes de recuperació d'informació, tant els mètodes teòrics com la indexació pràctica, i se'n resumeixen els resultats. També s'explica àmpliament en què consisteix el model vectorial de recuperació d'informació i, finalment, es parla de les tècniques de recuperació en la Web i la relació amb l'anomenat *falsejament d'índexs* (en anglès, *spamdexing*), que és la capacitat d'aconseguir ocupar les primeres posicions dels motors de cerca.

Paraules clau: falsejament d'índexs, indexació, model vectorial de recuperació d'informació, recuperació d'informació.

Abstract: ***Finding documents through words and links***

This paper focuses on the retrieval of words and links to find documents. It expounds information retrieval methods, both theoretical methods and practical indexing, and sum-

marises the results. A broad explanation is also provided as to what the vector model of information retrieval is, and the paper finally addresses Web retrieval techniques and the relationship with the so-called *spamdexing,* which is the capacity to occupy leading positions in the search engines.

KEY WORDS: spamdexing, indexing, vector model of information retrieval, information retrieval.

1. INTRODUCCIÓN

En la segunda mitad del siglo XX se produce lo que se ha dado en llamar *explosión documental*: un crecimiento exponencial de la masa de documentos, de todo tipo y en todo soporte. Esto ha puesto de relieve el problema de la recuperación de información. Es decir, la necesidad de seleccionar documentos concretos que resuelvan necesidades informativas concretas. El problema se centra fundamentalmente en seleccionar en función del contenido de los documentos; otro tipo de selección (por fechas, autores, etc.) ofrece menos problemas, al tratarse de información estructurada que puede procesarse mediante tecnología convencional (Van Rijsbergen, 1979). La vía clásica de abordar dicho problema de la recuperación de información es la indización manual: el contenido de los documentos es examinado y analizado por personas expertas, y descrito por éstas utilizando los llamados lenguajes documentales: una suerte de lenguajes artificiales controlados diseñados específicamente para describir el contenido temático de los documentos (las materias de éstos). El resultado de estas descripciones documentales puede ser almacenado de forma que se faciliten búsquedas posteriores entre estas descripciones, y seleccionar así los documentos que puedan responder a unas determinadas materias. En un principio esta forma de almacenamiento eran los clásicos ficheros en papel o cartulina, ordenados por diversos criterios; y, posteriormente, las bases de datos convencionales de los ordenadores. La indización manual, sin embargo, aun cuando se almacenen y gestionen sus resultados con ordenadores, tiene serios inconvenientes. En primer lugar, es un proceso caro y costoso: debe ser llevado a cabo por personal especializado y se trata de una tarea que requiere tiempo; no se trata, pues, de una cuestión solamente de elevados costes económicos: el tiempo necesario para indizar los documentos es mayor que el que éstos tardan en producirse. Es imposible procesar ni siquiera una mínima parte de los documentos que se producen; el alto grado de obsolescencia de buena parte de la documentación actual agrava este problema. El segundo gran problema de la indización manual es el de la inconsistencia. Se ha comprobado experimentalmente que distintos indizadores describen el mismo documento de maneras distintas (a pesar de utilizar el mismo lenguaje controlado para ello) (Hooper, 1965; Stubbs *et al.*, 2000). Incluso el mismo indizador, en momentos diferentes, produ-

ce descripciones diferentes de los mismos documentos. Es difícil producir después una recuperación eficaz, partiendo de descripciones de contenidos inconsistentes: ¿qué materias se deberían buscar para satisfacer una determinada necesidad de información? Lo cual nos lleva al tercer problema: para seleccionar los documentos que resuelvan una necesidad de información, es preciso describir dicha necesidad, y hacerlo con el mismo lenguaje controlado que se utilizó para describir los documentos; si para esto fue necesario utilizar personal especializado, para formalizar las necesidades de información también será preciso. El usuario deberá recurrir a intermediarios, a ese personal especializado, para obtener resultados satisfactorios.

2. Métodos en la recuperación de información

En la actualidad, buena parte de los documentos están disponibles en formato electrónico. En ocasiones, documentos en soporte papel están también en formato electrónico, pues han sido elaborados mediante máquinas electrónicas (por ejemplo, con un procesador de texto); en otros casos, existen sola y directamente en soporte electrónico. Sea como fuere, este hecho introduce un cambio sustancial, pues, al estar el documento completo en un soporte legible por ordenador, puede ser procesado por programas informáticos y es posible plantearse una indización totalmente automática. La indización automática, sin embargo, no está exenta de problemas. El principal de ellos es que un documento contiene mucha información, pero débilmente estructurada; al menos, estructurada de una forma que no es lo suficientemente explícita como para que los programas informáticos actuales puedan entenderla. Una solución simple a este problema es lo que se ha venido conociendo como *búsquedas en texto libre*, o también como *búsquedas de subcadenas*. Esto es, la selección por parte de un programa informático de aquellos documentos que contienen tal o cual palabra. Normalmente se podrá buscar más de una palabra, y, en ese caso, se podrán indicar restricciones adicionales mediante operadores booleanos, operadores de proximidad, truncamientos, etc. Esta solución simple tiene sus inconvenientes: los más importantes son los derivados de la sinonimia y la polisemia. Dado que un mismo concepto puede expresarse con palabras distintas (sinónimos), no siempre se puede saber cuál de ellas habrá sido utilizada en cada documento; de otro lado, puesto que una misma palabra puede referirse a conceptos diferentes, podemos encontrarnos con que muchos documentos que contienen una determinada palabra en realidad tratan sobre temas que nada tienen que ver con lo que nos interesa. El uso de operadores booleanos, de proximidad, etc. puede ayudar, pero hace que las búsquedas sean difíciles de realizar por el usuario no experto, sin llegar a paliar, sin embargo, los problemas apuntados. En todo caso, las búsquedas por palabras contenidas en los documen-

tos producen un resultado en el cual todos los documentos encontrados lo son en la misma medida: no hay forma de saber qué documentos pueden ser mejores para satisfacer nuestra necesidad de información, y esto puede ser un problema cuando los documentos encontrados son muchos.

2.1. *Los modelos teóricos*

La superación o, al menos la mitigación de estos problemas, ha dado lugar a numerosos modelos teóricos; algunos de ellos no han sido aplicados nunca en la práctica. Otros, no obstante, son la base de los sistemas de recuperación más avanzados disponibles actualmente.

Un esquema de los principales modelos para la representación y búsqueda es el que se puede ver a continuación (figura 1), cuyas características desarrollamos a continuación:

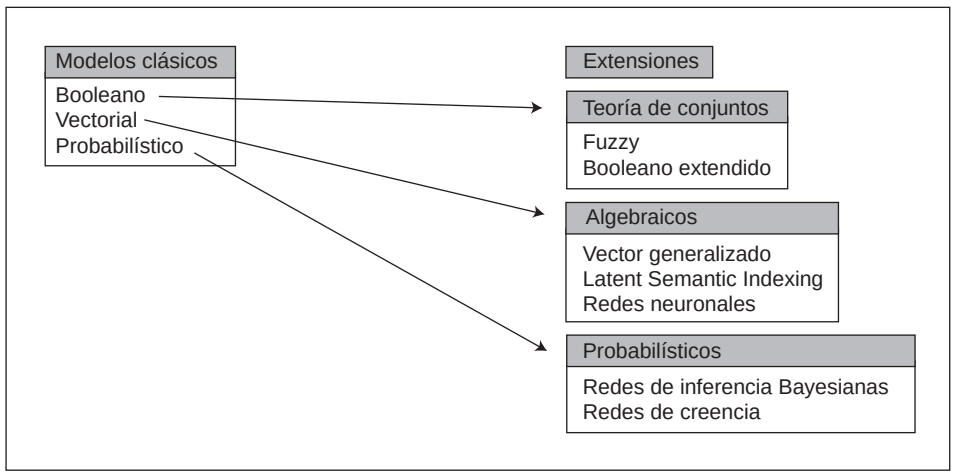


FIGURA 1. Modelos para la representación y búsqueda de palabras

a) Las características más importantes del modelo booleano son:

— Documentos.
  • Suele realizarse indización manual: a partir de la lectura y comprensión del texto, el indizador decide asignar los mejores términos que representen su contenido: descriptores.
— Consultas.
  • Las consultas se formulan utilizando los términos índice (descriptores) y una serie de operadores (booleanos, de proximidad, selección, truncamiento, etc.) y facilidades (índices, tesauros, etc.).

— El sistema de recuperación es sencillo, todo el esfuerzo recae en el usuario a la hora de plantear la consulta.
— Típico en bibliotecas, OPAC, etc.

b) En el modelo vectorial y probabilístico las características esenciales son:
— Documentos.
• Se lleva a cabo una indización automática: proceso complejo que trata de asignar automáticamente los mejores términos índice a los documentos (selección y extracción de términos).
— Consultas.
• Las consultas se realizan en lenguaje natural.
• El mismo proceso de indización automática se aplica a la consulta para obtener los términos índice que la representan.
— El sistema de recuperación es complejo. Todo el esfuerzo recae en él.
— Típico en motores de búsqueda de Internet (los mejores motores añaden información de enlaces, ej. Google).

2.2. *Indización manual vs. indización automática*

En el proceso de indización lo que pretendemos es obtener un conjunto de términos o procedimientos sintácticos (frases nominales) y convencionales para representar el contenido de un documento, con el fin de permitir su recuperación. Para ello nos basamos en el concepto de *término índice*: palabra o conjunto de palabras que tiene significado propio y que se utiliza para representar un concepto y en la idea de que tanto los documentos como la necesidad informativa pueden representarse utilizando términos índice.

Podríamos decir que *la indización* es el proceso de análisis que obtiene la representación de un documento / necesidad informativa utilizando términos índice.

Las características tipológicas de indización són las siguientes:

a) En el caso de la indización manual las características más importantes serían:
— Indización: conjunto de términos o procedimientos sintácticos (frases nominales) y convencionales para representar el contenido de un documento, con el fin de permitir su recuperación.
• Muy costosa en tiempo: muy lenta, mucho más que la producción de documentos.
• Muy costosa en dinero.

— Problemas de inconsistencia inevitable entre indizadores (sinonimia, polisemia, etc.), se requieren índices de concordancia y control de autoridades.
  • Dos personas pueden asignar diferentes palabras al mismo concepto, y la misma palabra puede aparecer en documentos que traten temas diferentes:

vendo coche usado vs. *automóvil de segunda mano*

b) En el caso de la indización automática algunas de sus características serían:
— Proceso complejo que asigna automáticamente los mejores términos índice a los documentos.
— Se persigue que las consultas puedan realizarse en lenguaje natural (texto libre).
— Problemas:
  • Información pobremente estructurada.
  • Formatos de documentos.
  • Codificación de la información.
  • Problemas de detección y conversión.
  • Normalización de términos (mayúsculas/minúsculas, acentos…).

Los pasos fundamentales que es necesario dar serían los siguientes:

1) Análisis del texto para determinar el tratamiento que se realizará sobre números, guiones, signos de puntuación, tratamiento de mayúsculas y/o minúsculas, etcétera.

2) Eliminación de palabras vacías, muy frecuentes y muy poco frecuentes. Se reduce el número de términos con valores muy pocos significativos para la recuperación.

3) Aplicación de lematización sobre los términos resultantes para eliminar variaciones morfosintácticas y obtener lemas.

4) Selección de términos que serán considerados términos índice (sustantivos, nombres propios).

5) Utilización de tesauros. Puede ayudar tanto en el proceso de indización como en el de búsqueda de información (expansión de consultas).

Analicemos brevemente cada uno de estos pasos.

2.2.1. *Análisis del texto (tokenización)*

Los elementos a tener en cuenta en esta fase son:

— Separación de palabras y «localización».
  • Carácter espacio, punto, comas, etc.
— Caracteres de puntuación.
  • A veces forman parte de términos (TVE-1, sub'21, Canal+, *e-mail*).
— Tratamiento de acentos.
  • Importante en otras fases del proceso léxico.
— Tratamiento de números.
— Detección de sintagmas y grupos nominales.
  • Nombres propios y expresiones multipalabra.
— Almacenamiento en mayúsculas/minúsculas.

2.2.2. Palabras vacías, muy frecuentes y muy poco frecuentes *(stop word)*

Se pretende reducir el ruido que pueda introducir la indización de todos los términos de un documento, y esta reducción se consigue suprimiendo:

— Palabras vacías:
  • Poseen muy poca capacidad semántica.
— Palabras muy frecuentes:
  • Si un término aparece en casi todos los documentos no sirve para diferenciar unos de otros.
— Palabras muy poco frecuentes.
  • Suelen ser errores de tecleado o palabras muy específicas (la probabilidad de que un usuario las solicite es muy baja).

2.2.3. Proceso de lematización *(stemming)*

En el proceso de lematización se tienen en cuenta los aspectos siguientes:

— «En un diccionario o repertorio léxico, elegir convencionalmente una forma para remitir a ellas todas las que derivan de su misma familia por razones de economía» (DRAE, 22.ª ed.).
— Palabras que son variaciones morfológicas con un significado prácticamente idéntico.
— Tratamiento:
  • Simple: eliminación de plurales (*s-stemmer*) o sufijos.
  • Complejo: sofisticadas técnicas de análisis procedente del PLN.
— Se basan en:
  • Aplicación de reglas.
  • Autómatas finitos.

2.2.4. Selección de términos índice

Con el objetivo de reducir la carga computacional, se intentan seleccionar los mejores términos índice.

Posibilidades de la selección de términos índice:

a) Valor de discriminación: capacidad de un término para discriminar unos documentos de otros. Tiene un coste computacional muy elevado. Además está relacionado con la frecuencia de aparición del término en toda la colección de documentos.

b) Naturaleza morfosintáctica del término: las palabras que actúan como nombres tienen mayor contenido semántico. Se pueden emplear técnicas del PLN para esta tarea, pero su coste computacional es muy elevado en comparación con sus beneficios.

2.2.5. Aplicación de tesauros

Un tesauro es un diccionario de términos controlados que contiene relaciones entre términos.

Los usos en recuperación de información (RI) son los siguientes:

— Indización (generalmente manual):
  • Los tesauros proporcionan un vocabulario controlado para la normalización de conceptos.
— Consultas:
  • Los tesauros se utilizan para plasmar con mayor exactitud la necesidad informativa del usuario, o bien, para reducir o ampliar los resultados del sistema en función de la jerarquía de términos presentes en el tesauro.
  • Expansión de consultas: trata de plantear una nueva consulta añadiendo nuevos términos relacionados con los de la consulta original (es necesario realizar un recálculo de pesos).

3. La aproximación lingüística[1]

Durante la década de los noventa, la disciplina conocida como procesamiento del lenguaje natural (PLN) experimentó un fuerte impulso que posibilitó el desarrollo de técnicas de análisis robustas, es decir, aplicables a textos sin restricciones de dominio, lo que, a su vez, permitió ampliar sus campos de aplicación. Uno de los destacados es el de la recuperación de información (RI).

Desde el campo del PLN no tardó en observarse cómo el método de indexación comúnmente adoptado en RI era resultado de un análisis muy superficial del

1. Figuerola *et al.*, 2006.

texto, y que éste podía perfeccionarse empleando las nuevas herramientas de análisis desarrolladas, para solucionar o, cuando menos, atemperar los efectos que más se denunciaban en RI —y que aún padecemos hoy en día en nuestra búsqueda cotidiana en Internet como determinantes a la hora de aumentar la efectividad en los sistemas de recuperación de información: los derivados de la ambigüedad léxica, tanto en el ámbito de la categoría gramatical como en el de significado. Como se explicó en el apartado anterior, la representación de documentos y preguntas consistía —y consiste, aún hoy día, en la mayoría de los sistemas en uso— en la detección de las «palabras ortográficas» (al menos para las lenguas con nuestros convenios ortográficos) de los textos, la normalización de las mismas a su forma mayúscula y minúscula (con eliminación de acentos y diacríticos) y la supresión de las que están incluidas en lo que se conoce como *listas de parada* o *listas de palabras vacías*.

Independientemente del método de «pesado» adoptado y de la «función o métrica de comparación» de preguntas y documentos que cada sistema implemente —que determinará, como se ha dicho también en el apartado anterior, los documentos a recuperar y el orden en que se devuelven al usuario—, el conjunto inicial de documentos candidatos susceptibles de ser recuperados será seleccionado entre aquellos que contengan, dependiendo del sistema de recuperación, todas las mismas palabras de la consulta (caso, por ejemplo, de Google), o al menos una parte de las mismas palabras de dicha consulta (caso de los sistemas basados en el modelo vectorial).

Repasamos a continuación los diferentes experimentos que se han planteado sobre colecciones monolingües y que, siguiendo a Tzoukerman *et al.* (1997), pueden dividirse en propuestas en indexación morfológica, indexación sintáctica e indexación basada en el sentido de las palabras.

3.1. *Indización morfológica*

En RI se han propuesto y experimentado técnicas no lingüísticas para intentar indexar las palabras de los documentos y de las preguntas por su raíz (técnicas de *stemming*). Estos métodos no lingüísticos, sencillos y eficientes computacionalmente, simplemente realizan una poda indiscriminada de, normalmente, determinados fines de palabra.

Se han propuesto métodos que van desde un simple *s-stemmer*, es decir, aquél que, para el inglés, elimina de toda palabra el carácter final *s* (con lo que se busca que los plurales y singulares de las palabras de documentos y preguntas se indexen por un mismo patrón), hasta otros más sofisticados para intentar tratar la morfología derivativa. Obviamente estas eliminaciones ciegas de ciertos sufijos producen anomalías en el intento de obtención de la raíz tanto por exceso como por defecto. Una versión del conocido algoritmo de Porter normaliza a la forma *organ* las palabras *organization*, *oganism* y *organ* (Krovetz, 1993). Una versión de un *s-stem-*

mer para el español que elimina los sufijos *as*, *es*, *os*, *a*, *e* y *o* de todas las palabras tiene, por ejemplo, como efecto transformar tanto *capa*, *capo* (y versiones plurales) y *cape* en *cap* (Figuerola *et al.*, 2002).

Como quiera, además, que una misma palabra puede tener, para diferentes categorías gramaticales, también la misma forma canónica (por ejemplo, *bajo* es la misma forma canónica cuando es adjetivo, sustantivo y preposición), se ha de buscar una forma de representación, en el momento de la indexación, diferenciada (*bajo*/A, *bajo*/P, *bajo*/S). De este ejemplo que hemos puesto puede colegirse fácilmente que el efecto de la desambiguación categorial puede ser beneficioso, pues con el par canónica/categoría gramatical se discriminan diferentes usos (acepciones) de la cadena de caracteres *bajo*. Otros efectos positivos que pueden obtenerse de utilizar técnicas de *POS-Tagging* en la indexación son: una eliminación coherente de las palabras vacías (por ejemplo, desechar *bajo* como preposición como palabra de indexación) y una posibilidad de reducción del tamaño de los índices (Chowdhury y McCabe, 1998).

En cuanto a los resultados obtenidos en los distintos experimentos de indexación morfológica en el momento de la recuperación, lógicamente han sido dependientes del lenguaje de la colección documental, pues los diferentes fenómenos morfológicos (flexión, derivación y composición) no se manifiestan con la misma intensidad en todas las lenguas (el inglés, p. e., es un idioma muy pobre a nivel flexivo en comparación con el español; el alemán, por otro lado, es un idioma muy aglutinativo). Así, por ejemplo, para el inglés, la conclusión obtenida es que la indexación con técnicas lingüísticas no aporta mejoras respecto de los métodos no lingüísticos, con lo que no resulta aconsejable el uso de las primeras dado la diferencia en el coste computacional. Respecto del español, los resultados obtenidos en Figuerola *et al.* (2002) parecen indicar que las técnicas de *stemming* producen efectos beneficiosos frente a los métodos que no realizan ninguna normalización. Para otros idiomas, como por ejemplo el holandés y el alemán, se ha comprobado que tratar la descomposición de palabras ortográficas en las correspondientes gramaticales produce efectos beneficiosos, tanto utilizando técnicas lingüísticas (Kraaij y Pohlmann, 1998; Monz y Rijke, 2002) como no lingüísticas (McNamee y Mayfield, 2002). En cuanto a la evaluación de los efectos que pudieran derivarse de los errores en la desambiguación categorial (la precisión de los *POS-Taggers* se sitúa entre el 95 % y el 97 %, o incluso superior), según se desprende de Gonzalo *et al.* (2002), no parecen relevantes.

3.2. *Indización sintáctica*

El método de indexación por palabras aisladas implícitamente asume la independencia de éstas respecto de los textos de las que se extraen y, por tanto, obvia lo siguiente:

1) Muchos conceptos se construyen concatenando, en determinadas lenguas, varias palabras ortográficas. Ese conjunto de palabras puede tener, para determinados dominios semánticos, una gran relevancia y, sin embargo, aisladamente, ese conjunto de palabras, por ser muy utilizadas en la colección documental, adquirir un peso irrelevante. Además, el orden de las palabras en la frase implica una variación del significado (*college junior*, vs. *junior in college* vs. *junior college*).

2) Por otra parte, determinados conceptos pueden expresarse con diferentes construcciones sintácticas que sería conveniente, a la hora de indexar, buscar una representación común (*Poland is attacked by Germany* vs. *Germany attacks Poland*).

Las conclusiones obtenidas por los grupos de investigación que más han experimentado en la indexación de sintagmas (grupo Xerox, grupo Clarit y Strazalkowski *et al.*, fundamentalmente) con técnicas lingüísticas pueden resumirse en las siguientes: en la indexación por sintagmas aunque se obtienen mejores resultados utilizando técnicas lingüísticas que meramente estadísticas, las diferencias son escasas; las mejoras entre una indexación por sintagmas con técnicas lingüísticas y una indexación por simples palabras ortográficas son inapreciables si las preguntas son cortas, aunque si las preguntas son largas sí se aprecian; la indexación por sintagmas no debe suplir a la indexación de los elementos simples que los componen; no es fácil determinar qué peso dar a los compuestos detectados.

3.3. *Indización basada en el sentido de las palabras*

Se han propuesto varios métodos para indexar documentos y preguntas de acuerdo al significado de las palabras que los componen, con el objetivo de medir los efectos que pudieran producirse al resolver los problemas de la ambigüedad léxica semántica. Para ello, se han utilizado diferentes recursos, principalmente los diccionarios y la red semántica de palabras WordNet (Peñas, 2004). La indización basada en los sentidos de acuerdo a un diccionario, dada su forma de organización, permite la representación diferenciada de los diferentes significados de un mismo significante. Esto es, posibilita el tratamiento de la polisemia y la homonimia. Utilizando una red semántica como WordNet, organizada en *synsets* (conceptos), es posible el tratamiento no sólo de los fenómenos anteriores sino también el de la sinonimia, además de la meronimia, hiponimia… dado que en la base de datos también se almacenan dichas relaciones entre los *synsets*. En cuanto a los experimentos aplicados a la indexación, resumiendo, se han concentrado en dos aspectos principales (Gonzalo *et al.*, 1999):

1) Evaluar si producen mejoras y en qué medida en la recuperación de información.

2) Fijar el umbral de error en la precisión de la desambiguación a partir del cual se produce una degradación en la efectividad de la recuperación de información.

De los resultados obtenidos del primer tipo de experimentos, los primeros efectuados cronológicamente, no era posible establecer unas conclusiones, dadas las tasas de precisión de los desambiguadores utilizados. Efectivamente, no se podía determinar si era beneficiosa o no en RI la indexación por sentidos, pues no era posible establecer la degradación que producía la desambiguación incorrecta. Otros experimentos han utilizado la estrategia de la desambiguación manual, pero para ello han recurrido a textos muy breves, p. e., pies de página (Smeaton y Quigley, 1996), con lo que los resultados no pueden extrapolarse a colecciones de grandes volúmenes de texto. El problema parece aún abierto, aunque más bien se ha pospuesto hasta que la tecnología en desambiguación madure. Independientemente de estos problemas enunciados, también se ha planteado el de la «granularidad» de los sentidos tanto en diccionarios como en WordNet. Un «grano muy fino» (trabajar con muchas acepciones diferentes para una entrada léxica), puede ser, muchas veces, contraproducente en RI, dado que al indexar separamos sentidos que pueden estar semánticamente muy cercanos.

3.4. *Expansión de consultas*

Uno de los problemas más importantes en RI consiste en formular la consulta para que plasme adecuadamente la necesidad informativa del usuario. Aparte de los requerimientos del sistema para formalizar la consulta, el mayor problema consiste en determinar el conjunto de palabras que expresen semánticamente esa necesidad. El problema se agrava debido al efecto de inconsistencia en la asignación subjetiva de términos a conceptos. Figuras como la sinonimia o la polisemia (u otras menos importantes, como la homonimia, la antonimia, la hiperonimia, la hiponimia, o la anáfora) hacen que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos. En esta situación no es de extrañar que el usuario tenga que replantear su consulta para obtener mejores resultados. De hecho, es ésta una de las acciones más habituales de los usuarios que utilizan motores de búsqueda en Internet. Se han propuesto diversos mecanismos para construir la nueva consulta. En general, en todos ellos se realiza una ampliación de nuevos términos a la consulta inicial y un recálculo de la importancia de cada término en la nueva consulta. Esto es lo que se conoce como expansión de consultas. Se pretende ampliar el número de términos que mejor definan la necesidad informativa del usuario de acuerdo a la colección documental y al modelo de recuperación utilizado. Para realizar la expansión lo más rápido sería utilizar tesauros o diccionarios generales ya existentes. Podemos realizar una clasificación de técnicas de expansión dependiendo de si requieren o no de la presencia del usuario. Según este punto de vista se distinguen dos grandes enfoques:

a) Realimentación de consultas utilizando criterios de relevancia del usuario (*user relevance feedback*). Requiere una buena interfaz con el usuario, pero es el mecanismo que mejores resultados proporciona. También se utiliza en motores de búsqueda en Internet, con la opción «páginas similares» o «*more like this*».

b) Expansión automática de consultas. No requieren de la presencia del usuario. Se pueden dividir a su vez en dos tipos:

— Análisis local. La expansión utiliza exclusivamente información de los documentos recuperados con la consulta inicial. Destacamos, por sus buenos resultados, la denominada pseudo-realimentación de consultas (*pseudo relevance feedback*). También se utilizan técnicas de agrupamiento local (tesauros locales de términos).

— Análisis global. Utiliza información de toda la colección de documentos para expandir la consulta. Se suelen emplear mecanismos de agrupamiento global con el objetivo de crear tesauros de términos. Destacamos varias técnicas: tesauros construidos a partir de la medida simple de coocurrencias, tesauros de similitud construidos realizando la transposición de la matriz documentos-términos (Qiu y Frei, 1993), tesauros construidos a partir de la asociación de términos y frases (*phrasefinder*), y tesauros basados en información sintáctica.

3.5. *Resumen de los resultados experimentales*

a) Aplicar lematización. Mejoras de 11,46 % y 10,85 % ($\bar{p}$ y *P*@10).

b) Realimentación de consultas con relevancia del usuario. El usuario visualiza los resultados y marca los relevantes y no relevantes y se reelabora la consulta. Hay mejoras del 300,1 % y 301,2 % ($\bar{p}$ y *P*@10).

c) Pseudo-realimentación de consultas. De forma automática se consideran los primeros documentos recuperados como relevantes. Algunas consultas mejoran y otras empeoran. Considerando los 5 primeros documentos recuperados y con 40 términos de más peso tenemos mejoras del 10,73 % y 8,43 % ($\bar{p}$ y *P*@10).

d) Tesauros. Las relaciones se pueden calcular automáticamente computando relaciones de coocurrencia tanto de términos como de documentos (tesauros de asociación); o si dos documentos poseen términos comunes (tesauro de similitud).

Además:

— Podemos utilizar tesauros globales (toda la colección) o locales (sólo los documentos recuperados).

— Podemos utilizar tesauros globales (toda la colección) o locales (sólo los documentos recuperados);

— los tesauros de asociación y los de similitud obtienen resultados similares, pero los de similitud tienen un tiempo de cómputo elevado;

— la expansión es mejor cuando se consideran los mejores términos relacionados con todos los términos de la consulta original;

— el empleo de tesauros locales obtiene mejores resultados.

4. El modelo vectorial

El modelo vectorial fue definido por Salton (Salton, 1968) hace ya bastantes años, y es ampliamente usado en operaciones de RI, así como también en operaciones de categorización automática, filtrado de información, etc. En el modelo vectorial se intenta recoger la relación de cada documento D_i, de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características.

En el modelo vectorial:

— cada documento es representado por un vector de términos;

— las consultas, formuladas en lenguaje natural, son representadas también como un vector de términos;

— es fácil aplicar alguna función de similitud que estime la semejanza entre el vector de la consulta y el de cada uno de los documentos.

Planteemos el problema de una manera más formal:

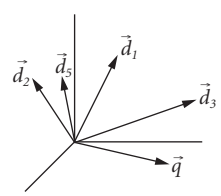
— cada documento d_j de la colección de N documentos se representa por un vector de m componentes, siendo m el número de términos índice de la colección;

— la consulta q se plantea al sistema en lenguaje natural, y también se representa por un vector;

— cada elemento del vector expresa la importancia que tiene el término índice en el documento o en la consulta: peso.

$$d_j \rightarrow \vec{d}_j = (w_{1j}, w_{2j}, \ldots, w_{mj})$$

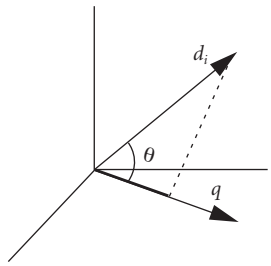
$$q \rightarrow \vec{q} = (w_{1q}, w_{2q}, \ldots, w_{mq})$$



— Para calcular la similitud entre documentos y consultas se supone que la distancia semántica entre ellos coincide con la distancia entre los vectores que las representan;

— normalmente esa distancia se mide por el coseno del ángulo que forman.

Los documentos se ordenan por orden de similitud con la consulta (ranking) y se presentan los primeros al usuario.



$$\frac{\sum_{i=1}^{m} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{m} w_{ij}^2 \cdot \sum_{i=1}^{m} w_{iq}^2}}$$

Podemos utilizar los denominados vectores binarios, para ello mostremos con un ejemplo su utilización.

Una colección de documentos en la que el total de términos distintos fuese $n = 4$.

TABLA 1. *Matriz de documento por término*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$
Doc_1	0	1	1	0
Doc_2	1	0	1	0
Doc_3	1	1	0	1
Consulta	0	1	0	1

Cada vector tiene $n = 4$ elementos, uno por cada término posible. El valor de cada elemento es 0 o 1, dependiendo de si el término aparece o no en el documento. Cualquier consulta puede ser tratada de la misma forma.

Si aplicamos una función de similitud simple, como el producto entre los vectores de la consulta y de cada documento:

TABLA 2. *Matriz con función de similitud*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	
Doc_1	0	1	1	0	simil. = 1
Doc_2	1	0	1	0	simil. = 0
Doc_3	1	1	0	1	simil. = 3
Consulta	0	1	0	1	

Obtenemos una lista de los documentos similares a la consulta, ordenados por similitud.

El que más se ajusta a la consulta es Doc_3, seguido de Doc_1.

Pero no solamente podemos utilizar el vector binario, lo más interesante es poder utilizar pesos, de esta forma:

— podemos registrar más información, no solamente la aparición de términos en documentos;

— un término puede ser más significativo en un documento que otro;

— podemos asignar a cada término un peso en cada uno de los documentos, en función de su importancia en cada documento;

— ese peso se puede estimar de diversas formas (por su frecuencia de aparición, por el lugar o campo del documento en que aparece, etc.);

— podemos representarlo mediante un valor numérico.

Un ejemplo mediante el empleo de pesos sería el siguiente:

TABLA 3. *Matriz de pesos con función de similitud*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	
Doc_1	0	0,7	0,2	0	simil. = 0,35
Doc_2	0,5	0	0,6	0	simil. = 0
Doc_3	0,6	0,4	0	0,2	simil. = 0,26
Consulta	0	0,5	0	0,3	

El documento que más se ajusta a la consulta es Doc_1.

El cálculo de los pesos puede hacerse por tres factores:

1) Si un término se repite mucho en un documento debe ser muy representativo de su contenido.

Operación: contar el número de veces que aparece un término en un documento (*tf*).

2) Si un término aparece en casi todos los documentos no sirve para diferenciar unos de otros.

Operación: contar el número de veces que aparece el término en toda la colección documental (*idf*).

3) Efectos laterales de documentos largos (muchos términos) frente a documentos cortos (pocos términos):

Operación: aplicar un factor corrector de normalización que es necesario porque:

— no todos los documentos tienen el mismo tamaño;

— conviene normalizar los pesos obtenidos con la frecuencia y el *idf*;
— el peso de un término *t* en un documento *d* se obtiene con estos tres elementos.

$$\frac{tf \times idf}{\text{normalización}}$$

Para poder trabajar con estos planteamientos se diseñaron diferentes sistemas de pesado, de forma que:
— se han propuesto diferentes formas de calcular cada uno de los tres componentes;
— cada una de esas formas se denomina o representa mediante una letra;
— las combinaciones posibles se denominan esquemas de pesado;
— ejemplo: BNN, NTC, ATU.

Para el cálculo de la frecuencia las formas son (en negrita la letra que se aplica al esquema):

none n_{tD}
binary 1
max-norm $\dfrac{n_{tD}}{\text{máx}\, n_D}$
aug-norm $0{,}5 + 0{,}5 \left(\dfrac{tf}{\text{máx}\ n_D} \right)$
square n_{tD}^2
log $ln(n_{tD}) + 1{,}0$

Donde:

n_{tD} n.º de veces que el término *t* aparece en el documento *D*
máx n_D n.º de veces del término que más aparece en el documento *D*

Para el cálculo del *idf* las formas son (en negrita la letra que se aplica al esquema):

none 1
tfidf $\log\left(\dfrac{N}{nd_t} \right)$

prob $\log\left(\frac{N - nd_t}{nd_t}\right)$

freq $\frac{1}{N}$

squared $\log\left(\frac{N}{nd_t}\right)^2$

Donde:

N número de documentos en la colección
nd_t número de documentos en que aparece el término t

Para el cálculo del normalizador las formas son (en negrita la letra que se aplica al esquema):

none 1

sum $\sum_{i=1}^{n} peso_{t_i D}$

cosine $\sqrt{\sum_{i=1}^{n} peso^2_{t_i D}}$

fourth $\sum_{i=1}^{n} peso^4_{t_i D}$

max máx $peso_{tD}$

Donde:

$peso_{tiD}$ peso del término i en el documento D
n número de términos en el documento D
máx $peso_{tD}$ peso del término con más peso en el documento D

Por ejemplo, si el esquema seleccionado fuera ntc-ntc (esquema en el proceso de indización y en el de consulta, que puede ser distinto), el cálculo sería:

Peso. Esquema ntc-ntc $\frac{tf \times idf}{\text{normalización}}$

— *tf* (*term frecuency*): número de veces que aparece un término en el documento/consulta.

— *idf* (*inverse document frecuency*): $\log\left(\frac{N}{nd_t}\right)$

N número de documentos en la colección
nd_t número de documentos en que aparece el término t

— normalización: se consigue haciendo que los vectores sean unitarios.

5. La recuperación en la Web

Las técnicas de recuperación de información que se han empleado en la Web, han procedido en su mayor parte de los sistemas de RI tradicionales. Por ello han surgido grandes problemas, debido a que el entorno de trabajo no es exactamente el mismo y además las características de los datos almacenados difieren considerablemente. Además han surgido nuevos problemas como el *spamming* o el enorme tamaño que deben soportar los índices, haciendo más difícil su adecuada gestión mediante el empleo de los modelos tradicionales. Las páginas web poseen una característica que las hace especiales. Prescindiendo de imágenes, sonido, elementos de captación de datos y demás ornamentos, las páginas web tienen enlaces con otras páginas. Estos enlaces son los que confieren su particular carácter a la documentación web (Alonso Berrocal *et al.*, 2003).

A partir de esos enlaces el espacio Web puede ser considerado como un grado dirigido, en el que los nodos serían las diferentes páginas existentes y los arcos, los hipervínculos que enlazan un nodo con otro (Alonso Berrocal *et al.*, 2004).

La explotación de la estructura hipertextual (Alonso Berrocal *et al.*, 1999) como método de recuperación incluye los lenguajes de consulta a la Web y la búsqueda dinámica, ideas que no están aún suficientemente implantadas. Los lenguajes de consulta a la Web pueden utilizarse para localizar todas las páginas web que tengan al menos una imagen y que sean accesibles al menos desde otras tres páginas, empleando para ello diversos modelos.

Este tipo de planteamientos se extrapola a la Web, considerado como una colección de documentos y por lo tanto se le aplican los modelos comentados. Pero le añadimos el matiz que nos suministran los enlaces, dándole un contenido semántico que podemos emplear en el modelo vectorial (Figuerola *et al.*, 2000).

Los trabajos más interesantes con enlaces están seguramente en el empleo de técnicas de posicionamiento.

5.1. *Técnicas de posicionamiento*

Las técnicas de posicionamiento, las podemos entender como el conjunto de procedimientos que permiten colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda. Estas técnicas han tenido y tienen un campo de trabajo y estudio muy activo y en el que se trabaja de forma constante.

Existen dos grandes variantes en los algoritmos de ranking:

— variantes del modelo vectorial o booleano

— los que siguen el principio de extensión de los enlaces.

De la primera variante hay tres métodos:

— booleano extendido

— vectorial extendido

— más citado.

De la segunda variante hay tres métodos:

— WebQuery

— HITS

— PageRank.

Algunas de las técnicas más utilizadas han sido las siguientes.

5.1.1. HITS

Este algoritmo desarrollado por Kleinberg (Kleinberg, 1999) depende de la consulta y considera el conjunto de páginas *S* que *apuntan a* o *son apuntadas por* la respuesta:

— páginas que tienen muchos links que apuntan a ellas en *S* son $\boldsymbol{A}(_{P})$= llamadas autoridades (*authorities*);

— páginas que tienen muchos links de salida son llamadas conectores $h(p)$ = conectores (*hubs*).

Mejores páginas *authorities* vienen de links de entrada desde buenos conectores (*hubs*) y buenos *hubs* vienen de enlaces de salida de buenas *authorities*.

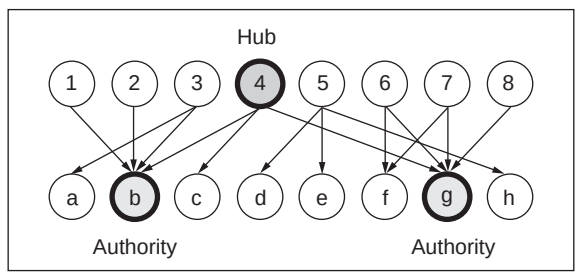


FIGURA 2. Ilustración del algoritmo HITS

5.1.2. PageRank

El PageRank (Page *et al.*, 1998) es la técnica de posicionamiento de mayor éxito y aunque se han descrito diversos problemas en el mecanismo básico de obtención del PageRank, se han planteado soluciones a los mismos (Sung Jin y Sang Ho, 2002) y constantemente se publican artículos sobre la mejora del mismo. La técnica del PageRank ha demostrado suficientemente sus características como técnica de posicionamiento en los procesos de recuperación de información (Dominich y Skrop, 2005).

El PageRank simula un usuario que navega aleatoriamente en la Web, quien salta a una página aleatoria con probabilidad q o que sigue un hyperlink aleatorio (en la página actual) con probabilidad $1 - q$.

Este proceso se modela como una cadena de Markov, en que se puede calcular la probabilidad estacionaria de estar en cada página.

La importancia de una página viene dada por la importancia de las páginas que la enlazan.

$$PR(a) = q + (1 - q) \sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

6. Webspam

Un campo de trabajo de gran actualidad son las investigaciones sobre w*eb spam*. No podemos decir con certeza que exista una única definición para *web spamming*, referido por muchos autores (Gyongyi y Garcia-Molina, 2005) como s*pamdexing*, y muchas veces definido como una práctica para conseguir una posición elevada en los resultados de los motores de búsqueda, utilizando técnicas para engañar a los algoritmos de clasificación.

El término *spam* según Castillo *et al.* (2006) ha sido utilizado en los últimos años referido a los mensajes no solicitados (normalmente comerciales).

El *spamdexing* es definido por Gyongyi y Garcia-Molina (2005) y referido por Castillo *et al.* (2006), como «cualquier acción con la intención de conseguir un aumento injustificado de la relevancia o importancia de una página web, considerando su valor real».

Cualquiera que sea la definición es cierto que el *spam* se refiere a algo indeseable, incluso perturbador, con una influencia negativa en el proceso HTTP, que al basarse en el paradigma solicitud-respuesta imposibilita el envío directo de las páginas por los *spammers* hacia los usuarios finales. Para superar esta defensa del

protocolo los *spammers* utilizan otras técnicas y medios. La más utilizada es a través de mensajes, aparentemente unidireccionales, vía *e-mail*.

Pero si nos centramos en el modo de operar de los *spammers* sobre los sistemas de recuperación de información en la Web, veremos que es diferente del resto. En este caso los principales destinatarios son los motores de búsqueda y la forma de engañar y minar las relaciones de confianza establecidas entre los usuarios de los motores de búsqueda (Gyongyi y Garcia Molina, 2005).

Estas técnicas de *spam* orientadas a los motores de búsqueda, pretenden obtener la atención de los usuarios finales, con fines normalmente comerciales. Una de las razones que subyacen están en las dificultades de los usuarios finales en distinguir las informaciones fiables de las no fiables debido al éxito de los motores en las últimas décadas (Metaxas y DeStefano, 2005).

Los usuarios han ido aumentando su confianza en los motores de búsqueda como medio de obtención de información, y los *spamers* han logrado, con éxito, conducir esa confianza a los resultados de cada consulta.

Para que sea posible continuar con la confianza en los resultados de las consultas, los constructores de motores de búsqueda deben realizar un gran esfuerzo para proporcionar respuestas sin *spam*. Realizarán sofisticadas estrategias de ranking que, junto a los algoritmos que permitan la detección del *spam,* lo eliminarán de las respuestas (Becchetti *et al.*, 2008).

De forma general algunas de las formas de realizar web *spam* se resumiría en la siguiente lista:

— *Spamdexing*
  • *keyword stuffing* (relleno)
  • *link farms* (granjas)
  • *spam blogs* (*splogs*)
  • *cloaking*.

6.1. SEO vs. *spam*

La optimización para motores de búsqueda (SEO, por sus siglas en inglés) tiene que ver con asegurarse de que un sitio sea encontrable por los buscadores. Los servicios que ofrecen los *spammers* incluyen la creación de miles o millones de páginas falsas que tienen como propósito el engañar a las máquinas de búsqueda y a sus usuarios.

En cualquier caso, la relación entre el administrador de un sitio web que intenta tener un alto posicionamiento y el administrador de la máquina de búsqueda es una relación entre adversarios en un juego de suma cero. Cada ganancia inmerecida de ranking para una página es una pérdida de precisión para la máquina de búsqueda.

Técnicas SEO legítimas (≈ técnicas de sombrero blanco):

— objetivo, aparecer en lo más alto cuando un cliente está buscándolos;

— en contraposición a una página elaborada por personas que odian a su cliente;

— más eficaz, pregunta a los sitios web legítimos para vincularse al cliente.

Spam (≈ técnicas de sombrero negro): crear lotes artificiales de los sitios web que enlazan a una página que promueve un producto (p. e. Viagra).

La separación en el «sombrero blanco» y el «sombrero negro» es una línea muy delgada.

7. Bibliografía

Alonso Berrocal; J. L.; Figuerola, C. G; Zazo, A. F. (2004). *Cibermetría*: *nuevas técnicas de estudio aplicables al Web.* Gijón: Trea.

— (1999). «Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información». *Scire*, vol. 5, n. 2, p. 91-98.

Alonso Berrocal, J. L. [*et al.*] (2003). «Agentes inteligentes: recuperación autónoma de información en la WEB». *Revista Española de Documentación Científica*, vol: 26, n. 1, p. 11-20.

Becchetti, L. [*et al.*] (2008). «Link analysis for web spam detection». *ACM Transactions on the Web,* vol. 2, n. 1, p. 1-42.

Castillo, C. (2006). A reference collection for web spam. *SIGIR Forum, vol. 40, núm. 2.*

Chowdhury, A.; McCabe, M. (1998). *Improving information retrieval using part of speech tagging* [en línia]. <*citeseer.ist.psu.edu/256084.html*> [Consulta: 29 mayo 2009].

Dominich, S.; Skrop, A. (2005). «Pagerank and interaction information retrieval». *Journal of the American Society for Information Science and Technology*, vol. 56, n. 1, p. 63-69.

Figuerola, C. G.; Alonso Berrocal, J. L.; Zazo Rodríguez, A. F. (2000). «El contenido semántico de los enlaces de las páginas web desde el punto de vista de la recuperación de información». A: Cabré, M. T.; Codina, L; Estopà, R (ed.). *Terminologia y Documentació.* I Jornada de Terminologia y Documentació (Barcelona, maig 2000). Barcelona: Institut Universitari de Lingüística Aplicada, 2000, p. 71-79.

Figuerola, C. G. [*et al.*] (2002). «Spanish monolingual track: The impact of stemming on retrieval». A: *Evaluation of Cross-Language Information Retrieval Systems.* Second Workshop of the Cross-Languge Evaluation Forum (Darmstadt, setembre 2001), Springer, vol. 2406, p. 253-261.

— (2006). «Encontrar documentos a través de las palabras». A: Fuentes, T.; Torres, J. (ed.). *Nuestras Palabras*: *Entre el Léxico y la Traducción*. Lingüística Iberoamericana, p. 147-174.

Gonzalo, J.; Peñas, A.; Verdejo, F. (1999). *Lexical ambiguity and information retrieval revisited*. 1999 Joint SIGDAT Conference on EMNLP and VLC (Maryland*, 1999),* p. 195-202.

GONZALO, J.; PEÑAS, A.; VERDEJO, F. (2000). *La indexación con técnicas lingüísticas en el modelo clásico de recuperación de información*. A: SANCHÍS, E.; MORENO, L.; GIL, I. (ed.). Primeras Jornadas de Tratamiento y Recuperación de Información (València, 4-5 juliol 2002). València: Universitat Politècnica de València. Facultat d'Informàtica, p. 97-106.

GYONGYI, Z.; GARCIA MOLINA, H. (2005). *Web spam taxonomy*. First International Workshop on Adversarial Information Retrieval on the Web.

HOOPER, R. S. (1965). *Indexer Consistency Test - Origin, Measurements, Results and Utilization.* Bethesda: MD.

KLEINBERG, J. M. (1999). «Authoritative sources in a hyperlinked environment». *Journal of the ACM*, p. 668-677.

KRAAIJ, W.; POHLMANN, R. (1998). *Comparing the effect of syntactic vs. statistical phrase index strategies for dutch*. Proceedings of ECDL'98 (setembre 1998), p. 605-617.

KROVETZ, R (1993). *Viewing morphology as an inference process.* A: KORFHAGE, R.; RASMUSSEN, E. M.; WILLET, P. (ed.). 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, 27 junio - 27 julio 1993). ACM, p. 191-203.

MCNAMEE, P.; MAYFIELD, J. (2002). «Language-Independent Approach to European Text-retrieval». A: *Cross-Language Information Retrieval Systems*. Springer, p. 29-139.

METAXAS, P. T.; DESTEFANO, J. (2005). «Web spam, propaganda and trust». AIRWeb2005, (10-14 maig).

MONZ, C.; RIJKE, M. (2002). «Shallow Morphological Analysis in Monolingual Information retrieval for Dutch, German and Italian». A: *Cross-Language Information Retrieval Systems.* Springer, p. 262-277.

PAGE, L. [*et al.*] (1998). *The PageRank citation ranking: Bringing order to the web* [informe técnico]. Stanford Digital Library Technologies Project.

PEÑAS, P. (2004) *Técnicas lingüísticas aplicadas a las búsqueda textual multilingüe: ambigüedad, variación terminológica y multilingüismo.* SEPLN.

QIU, Y.; FREI, H. P. (1993) *Concept-based query expansion.* A: KORFHAGE, R.; RASMUSSEN, E. M.; WILLET, P. (eds.). 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. (Pittsburgh, 27 junio - 27 julio 1993). ACM, p. 160-169.

RIJSBERGEN, C. J. VAN (1979). *Information Retrieval.* Glasgow: University of Glasgow. Department of Computer Science.

SALTON, G. (1968). *Automatic Information Organization and Retrieval.* Nova York: McGraw-Hill.

SMEATON, A.; QUIGLEY, I. (1996). *Experiments on using semantic distances between words in image caption retrieval*. A: FREI, H. P. [*et al.*] (ed.). 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Zuric, 18-22 agosto 1996). ACM, p. 174-180.

STUBBS, E. A.; MANGIATERRA, N. E.; MARTÍNEZ, A. (2000). «Internal quality audit of indexing: A new application of interindexer consistency». *Cataloguing & Classification Quaterly*, vol. 28, n. 4, p. 53-70.

SUNG JIN, K.; SANG HO, L. (2002). «An improved computation of the pagerank algorithm». *Lecture Notes in Computer Science*, vol. 2291. Springer, 2002, p. 73-85.

TZOUKERMAN, E.; KLAVANS, J.; JACQUEMIN, C. (1997). *Effective use of natural language processing of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsong*. Proceedings of 20th ACM/SIGIR (2 mayo 1997), p. 148-155.

Sessió I

Comunicacions

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 47-73 DOI: 10.2436/15.2503.02.2

Invitació a l'estudi estadístic del llenguatge

Rogelio Nazar
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

Resum

El tema d'aquesta presentació és la cruïlla interdisciplinària entre la lingüística i l'estadística. Està adreçada a lingüistes, per als quals pot tenir un interès teòric, o a professionals que treballen amb la llengua, per als quals pot tenir un interès pràctic. Enfoca el concepte de probabilitat de combinatòria de paraules des de tres perspectives diferents: *a*) els estudis d'associació entre les unitats que es combinen, *b*) la distribució en el corpus d'aquesta combinació d'unitats, i, finalment, *c*) les maneres de mesurar la similitud entre unitats d'acord amb les seves possibilitats de combinació. Tots aquests temes hi són tractats d'una manera estrictament teòrica i van acompanyats d'exemples d'aplicació pràctica en terminologia i en documentació. L'objectiu és demostrar que la utilització d'eines estadístiques en aquests camps és un complement necessari per a la intuïció dels investigadors.

Paraules clau: corpus textuals, estadística, lingüística quantitativa, llenguatge, probabilitat combinatòria.

Abstract: *Invitation to the statistical study of language*

The topic of this presentation is the interdisciplinary nexus between linguistics and statistics. It targets linguists, for whom it may have a theoretical interest, or professionals that work with language, for whom it may have a practical interest. It focuses on the concept of the combinatory probability of words from three different perspectives: *a*) the studies of association between the units that are combined, *b*) the distribution of this combination of units in the corpus, and finally *c*) the ways of measuring similarity between units according to the combination possibilities. All these topics are addressed in a strictly theoretical fashion and are illustrated by examples of practical application in terminology and in documentation. The objective is to demonstrate that the use of statistical tools in these fields is a necessary complement to the researcher's intuition.

KEY WORDS: text corpora, statistics, quantitative linguistics, language, combinatorial probability.

1. INTRODUCCIÓ

Aquesta comunicació està dirigida a persones que no tenen coneixements previs sobre l'encreuament interdisciplinari entre l'estadística i el llenguatge. Té el doble propòsit de ser una aportació des del punt de vista de la lingüística teòrica i, a la vegada, ser útil per a la documentació i per a la terminologia. Conseqüentment, inclou exemples de com aquest coneixement teòric es pot aplicar a la solució de problemes pràctics.

La intenció és introduir a la temàtica, però també conscienciar i aclarir. Conscienciar, perquè la lingüística quantitativa no només és una àrea marginal en lingüística, sinó que, a més, moltes vegades tant lingüistes com estadístics n'ignoren l'existència. Aclarir, perquè la relació entre estadística i llengua no és cap novetat ni pertany al món de les «noves tecnologies». Estem parlant d'una tradició que fa més de cinquanta anys que difon conceptes i mètodes que no tenen una relació inherent amb la informàtica. La utilització d'ordinadors és evidentment necessària per a dur a terme estudis en lingüística quantitativa, però parlar d'aquests temes no significa parlar d'un programa informàtic, perquè això equival a confondre el fenomen observat amb l'instrument d'observació. Certament, els mitjans són determinants, ja que, com deia Saussure, el punt de vista defineix l'objecte. Tanmateix, això no ha de dur a l'error de reificar les idees en la forma d'un programari. En definitiva, l'important és conèixer quins estudis s'han fet o es poden fer i prendre consciència que aquesta disciplina no es limita al recompte de vegades que dues paraules apareixen juntes en un corpus.

Pel que fa a la meva legitimitat com a orador, sóc aquí per la meva funció a l'IULA,[1] consistent a assimilar el coneixement que ja existeix sobre lingüística quantitativa, aplicar aquest coneixement a la solució de problemes pràctics i, a la vegada, intentar proposar algun coneixement nou en els fòrums científics.

No presento res de nou en aquesta comunicació. Faré, en canvi, un recorregut per algunes idees que he tractat ja en altres treballs. És important advertir que no represento necessàriament l'opinió dels meus companys de feina. Em refereixo particularment a un protocol que inclou un compromís amb la independència de llengua, que consisteix a esbrinar primer, i sempre que sigui possible, fins a quin punt es pot arribar a treure conclusions útils sense introduir coneixement explícit sobre una llengua en particular.

1. Institut Universitari de Lingüística Aplicada (*http://www.iula.upf.edu*).

Aquesta comunicació està organitzada de la manera següent: en la pròxima secció 2, analitzarem la confrontació existent entre dues formes molt diferents d'apropar-se a l'estudi de la llengua, davant de les quals la lingüística es troba en una posició ambigua: el món humanístic, per anomenar-lo d'alguna manera encara que sembli lleugerament imprecís, i el món científic, en particular el món de les «ciències dures» per oposició a les ciències socials, on el pensament quantitatiu és, a vegades, encara sospitós. A continuació, en la secció 3, entrarem en la matèria de l'anàlisi lingüística enfocada des de la perspectiva estadística. Analitzarem concretament el concepte de combinatòria de paraules. Veurem el concepte de probabilitat de combinatòria de paraules des de tres perspectives diferents: en la subsecció 3.1, els estudis d'associació entre les unitats que es combinen; en la subsecció 3.2, la manera en què aquesta combinació d'unitats es distribueix en un corpus i les conclusions que en podem derivar; i, finalment, en la subsecció 3.3, les formes de mesurar la similitud entre unitats d'acord amb les seves possibilitats de combinació. Com a exemple, analitzarem el bigrama i establirem el significat d'aquesta unitat més enllà de la seva definició formal, per saber en profunditat quin tipus d'informació codifica. Veurem que, encara que sembli sorprenent, la nostra identitat individual i col·lectiva està continguda en el bigrama. Com a exemple de les aplicacions pràctiques, en la secció 4 veurem la classificació de documents en diverses variants —subsecció 4.1 i 4.2—, així com elements per a la caracterització del significat i la desambiguació de terminologia i, en la subsecció 4.3, el descobriment de neologia. Existeixen altres possibilitats d'aplicació, entre les quals trobem línies de recerca en curs, com ara l'extracció automàtica de terminologia especialitzada o l'extracció de terminologia bilingüe de corpus no paral·lels, però aquestes línies, malgrat el seu interès, no es tractaran aquí per les limitacions d'espai.

2. El xoc entre dues cultures

Wilhelm Dilthey (1883) va advertir ja les diferències epistemològiques entre les ciències naturals, d'una banda, i les ciències socials i humanitats (o ciències de l'esperit), de l'altra, continuant una línia de pensament que va iniciar Kant. Mentre que en les ciències naturals preval un pensament mecanicista, amb el qual es pot predir la conseqüència de determinats esdeveniments, en les ciències de l'esperit, en canvi, aquest determinisme no és possible. La resposta d'un ésser humà davant d'un determinat esdeveniment és en última instància imprevisible. Fins i tot en aquestes circumstàncies, les ciències de l'esperit ens permeten almenys comprendre (*verstehen*) les circumstàncies històriques i individuals que envolten el que és humà.

La fita, però, en la història de la presa de consciència de la divisió de la cultura en el saber científic i el saber humanístic —divisió que encara estructura els cur-

rículums de l'educació secundària— és una conferència donada per C. P. Snow (1959), en la qual descriu la sospita mútua i la incomprensió existent entre científics i intel·lectuals. Tot i que pertanyin a les capes més educades de la població, els dos col·lectius són ignorants l'un de l'altre. Si bé després va moderar el seu discurs, en aquella ocasió Snow va plantejar que la gent que té un pensament de tipus tècnic és en general inculta, i els intel·lectuals, per la seva part, hostils a aquest pensament, són generalment incapaços de comprendre els conceptes científics més elementals.

Aquesta separació és particularment interessant en el si de les ciències socials, considerades «ciències toves» per oposició al rigor de les ciències naturals, les «ciències dures». La inclinació dels científics socials per una o per una altra branca de pensament dependrà de l'orientació ideològica personal o de la de cada facultat o departament, però entre els intel·lectuals de les ciències socials és comú advertir una reticència *a priori* cap a tot pensament de tipus tècnic en l'estudi del que és humà. Aquesta reticència està representada en la idea de Cornelius Castoriadis (1975) sobre el fet que amb un llenguatge reduït a allò que és instrumental es pot operar i calcular, però no es pot pensar, una idea amb ressonàncies a la polèmica constatació feta per Heidegger sobre la idea que «la ciència no pensa».

En sociologia, aquesta diferència va estar clarament representada per l'oposició entre el pensament crític i la reflexió filosòfica i històrica de l'Escola de Frankfurt davant l'hàbit dels sociòlegs nord-americans de la Mass Communication Research de promoure l'aplicació de mètodes quantitatius per sobre de la reflexió teòrica, enfrontament que va continuar tot i la col·laboració entre alguns dels màxims exponents d'ambdós bàndols, com Theodor Adorno i Paul Lazarsfeld.

El cas és particularment interessant en la lingüística, si es vol, «la més dura de les ciències toves». Fins i tot lingüistes experimentats expressen sorpresa en prendre consciència que existeix una lingüística quantitativa. Els que són «de lletres» no saben «de nombres». Mandelbrot (1961) encara estava en el moment oportú per a revitalitzar la pregunta sobre què és la lingüística i establir una diferència entre gramàtics i lingüistes. En el cas dels primers, preval el coneixement d'una llengua en particular i del que pot ser i el que no pot ser gramaticalment correcte; mentre que, segons aquest autor, la lingüística pertany al món de les ciències dures, i, en aquest sentit, l'important no són tant les característiques particulars, que són d'una infinita diversitat, sinó les propietats estructurals del llenguatge (actitud contra la qual Saussure segurament no tindria res a dir). L'estudi d'aquestes propietats possibilita enunciats científics amb una validesa que transcendeix el coneixement que es tingui d'una llengua en particular, la qual cosa està d'acord amb l'esperit científic que és procliu a la generalització, ja que no hi ha o no hi hauria d'haver ciència del que és particular.

L'encreuament interdisciplinari, però, és difícil. Les persones que venim d'àmbits més propers a la lingüística en general estem poc informats sobre els conceptes matemàtics més elementals i resulta laboriós començar de zero en el camp, sobretot per a qui no té els hàbits de pensament de les ciències dures. Tanmateix, aquest és, sens dubte, un camp d'estudi que justifica el desafiament; per això, per mitjà d'aquesta presentació, pretenc contagiar l'interès i aportar arguments a la confusió de les barreres entre ciències dures i toves, o entre coneixement científic i coneixement humanístic en general.

Aquestes barreres ja es confonen i la lingüística no n'és l'únic exemple. La teoria literària, món humanístic per antonomàsia, comença a patir també el setge de l'estadística. Un exemple n'és l'aportació que l'estadística està fent en les disputes sobre l'autoria d'obres literàries, en casos que inclouen figures prominents com la de Shakespeare (Vickers, 2002).

3. La informació com a probabilitat

En la línia de Shannon (1948) podem estimar la *quantitat d'informació* com la probabilitat d'ocurrència d'un signe en un missatge, una mesura de la quantitat de sorpresa que ens pot provocar un determinat esdeveniment. Per explicar-ho amb paraules senzilles, en determinats contextos sabem que hi ha esdeveniments que són més o menys normals i d'altres, inesperats. En el llenguatge hi ha certes concatenacions que són més predictibles que d'altres. Si cada dia, en sortir de la feina, el cap diu «fins demà» al treballador, després d'una sèrie d'esdeveniments d'aquest tipus l'enunciat resulta poc informatiu. Però si un determinat dia el text canvia per «aquesta empresa ja no seguirà comptant amb els seus serveis», direm que aquest segon enunciat és comparativament més informatiu, és a dir, causa major sorpresa. Aquesta sorpresa està directament relacionada amb la probabilitat d'aparició d'aquest missatge (la sorpresa no serà tan gran si l'empleat està acostumat a ser acomiadat de diferents feines).

El criteri de la freqüència com a estimació de probabilitat és el mateix que apliquem quan ens trobem en la situació de treure boles d'una urna. Si suposem que cada bola té la mateixa probabilitat de ser escollida, si en treure les boles d'una en una observem que les boles de vegades són negres i altres vegades són blanques, i després de treure cent boles ens adonem que hem obtingut norantacinc boles negres, aquesta circumstància, encara que sigui de manera intuïtiva, ens farà sospitar que la propera bola, la 101, tindrà un 95 % de probabilitats de ser negra.

Podem aplicar aquesta intuïció a l'estudi del llenguatge i adjudicar així un valor d'informació als signes, d'acord amb la seva probabilitat d'aparició en un missatge. En la fórmula [1], la probabilitat d'aparició del signe *i* és expressada com a

$p(i)$, $f(i)$ seria la freqüència d'una determinada paraula en un determinat corpus i N, la quantitat total de paraules d'aquest corpus.

$$p(i) = f(i) / N \quad [1]$$

En el lèxic tenim paraules que són més o menys informatives. L'aparició de paraules com *el*, *de* o *que* en un text ens sorprèn poc, i per això diem que són poc informatives. Si ordenem totes les paraules d'un corpus per freqüència decreixent, observarem que la freqüència d'una unitat està en funció de la seva posició en el rang (r); per tant, es compleix —aproximadament— la fórmula [2]:

$$f(x) = 1/r \quad [2]$$

Si multipliquem la freqüència d'una unitat pel seu rang (equació [3]) obtenim un valor constant c.

$$c = f \cdot r \quad [3]$$

La corba de la funció [2] representa també la distribució de la renda en les societats capitalistes —la llei de Pareto— per a Vilfredo Pareto, que la va descriure el 1906. Ordenats de major a menor renda, s'adverteix com són uns pocs els individus que posseixen la major part de la riquesa, mentre que la gran majoria en percep una mínima part. Entre els lingüistes, el descobriment s'atribueix a J. Estoup, l'any 1916, tot i que va ser divulgada per G. Zipf l'any 1949. L'interès per la llei de Zipf va decaure, però, a partir de l'estudi de Mandelbrot (1961), que la va reformular (fórmula [4]) per tal que s'adaptés millor a les dades observades, particularment en els rangs més alts i més baixos de la corba.

$$f(x) = P \cdot (r + p)^{-B} \quad [4]$$

En la fórmula de Mandelbrot, f és la freqüència i r el rang, mentre que P, p i B són paràmetres constants. Herdan (1964), però, objecta que aquests paràmetres no són constants sinó que depenen de la mida del corpus. La conseqüència d'això és que la fórmula no podria ser aplicada per a la comparació de mostres de mida diferent amb la finalitat, per exemple, de comparar la riquesa lèxica de les mostres.

La riquesa del vocabulari està directament relacionada amb la quantitat d'informació dels signes, la qual cosa determina el grau de dificultat de lectura o densitat d'un text. Això és el que Mandelbrot anomena la *temperatura del discurs*. En el seu cas, plantejava la relació entre l'extensió i el vocabulari d'un text, és a dir,

la quantitat de paraules diferents dividida per la quantitat total de paraules. Però podem establir diferents mesures de riquesa del vocabulari per a un autor o un text no solament segons això, sinó també posant en relació un text analitzat amb un coneixement previ que puguem tenir de la llengua en què està escrit. Aquest coneixement previ pot tenir la forma d'un model de llengua elaborat sobre la base d'un corpus d'una extensió de *n* milions de paraules, un corpus que podríem anomenar *corpus de referència* d'una llengua, conformat per textos de premsa o d'altres gèneres, que pertanyen a una determinada llengua o varietat dialectal. Mal anomenat «corpus de referència», perquè aquest corpus, per més gran que sigui, sempre tindrà un determinat biaix i no arribarà a ser veritablement una referència de la llengua. Aquest model, però, ens permet saber la raresa de les paraules que utilitza un text (o un autor), ja que per a nosaltres representaria un estàndard de llengua «normal».

3.1. *Associació*

Malgrat l'interès que pugui tenir l'assignació individual d'informació per als signes, és molt més interessant estimar les seves probabilitats de combinatòria. Si els signes es combinessin en el llenguatge de manera aleatòria, les seves probabilitats de combinació serien iguals a la multiplicació de les seves probabilitats individuals. La probabilitat de combinació aleatòria de les paraules *i* i *j* (fórmula [5]) defineix que la probabilitat d'aparició conjunta de *i* i *j* (expressada aquí com a intersecció) és igual a la de *i* multiplicada per la de *j*.

$$p(i \cap j) = p(i) \cdot p(j) \qquad [5]$$

Hi ha una aclaparadora quantitat i diversitat de mesures per a calcular les probabilitats de combinació de les paraules —o esdeveniments, en general— (Muller, 1973; Manning i Schütze, 1999; Evert, 2004; entre altres). En lingüística podem veure aquestes mesures aplicades a l'extracció de terminologia especialitzada polilexemàtica o a l'estudi de les col·locacions, tot un capítol en l'estudi del llenguatge. Les combinacions de paraules no són donades solament per la gramàtica, i això té indubtablement el seu correlat en les freqüències de coocurrència. En anglès, es diu *strong coffee*, però no *powerful coffee*. No obstant això, diem una *powerful computer*, però no una *strong computer.*[2] En cada llengua, i fins i tot en cada

2. Aquest últim exemple és interessant, perquè actualment ambdues seqüències de paraules tenen pràcticament la mateixa freqüència a Google; cosa que pot enganyar l'usuari desprevingut, perquè la segona forma, *strong computer*, apareix sempre formant part d'estructures més grans com *strong computer password.* És a dir, el nucli del que depèn *strong* no és en aquest cas *computer* sinó *password*, o *skills*, o *science background*, etcètera.

domini d'especialitat, existeixen certes preferències en les combinacions de paraules de diverses categories (verb-nom; adjectiu-nom; nom-nom, etc.). Per una raó pragmàtica, les coses acostumen a dir-se d'una determinada manera, i si bé la gramàtica ens permetria formular el text d'una altra, fent-ho així correríem el risc de confondre el receptor si ja existeix, en aquesta llengua, domini o registre, una manera típica o idiosincràtica de dir el que volem dir.

Les estadístiques d'associació ens poden informar sobre la manera típica en la qual es combinen les paraules d'una llengua perquè responen a la pregunta sobre quina és la probabilitat que dos esdeveniments ocorrin junts en una mateixa situació o, més precisament, si la freqüència d'aparició de dos esdeveniments en una mateixa situació es pot adjudicar a l'atzar. Un esdeveniment pot ser l'aparició d'una paraula i la situació pot ser un text, un paràgraf, una oració, una «finestra» de *n* paraules, etc. També es pot tractar de l'aparició de les paraules de forma concantenada, o no. Si es tracta d'una seqüència de dues paraules podem parlar d'un *bigrama*, d'un *trigrama* en el cas de tres unitats o d'un *n-grama* per a *n* unitats. Però cal tenir en compte que un *n*-grama podria ser definit d'una altra manera, com una seqüència de lletres o de categories morfològiques. A més a més, la coocurrència pot ser definida d'una manera diferent de la seqüencial. Podem definir *coocurrència* com l'aparició de les dues paraules en una finestra de context sense importarnos l'ordre en què apareixen. Les figures 1 i 2 mostren, per exemple, un criteri de coocurrència que consisteix a comprovar quantes vegades apareixen les paraules —a diferents distàncies i en diferent ordre— en una finestra de context de vint paraules.[3] En ambdós casos, estem analitzant les paraules que coocorren amb la forma

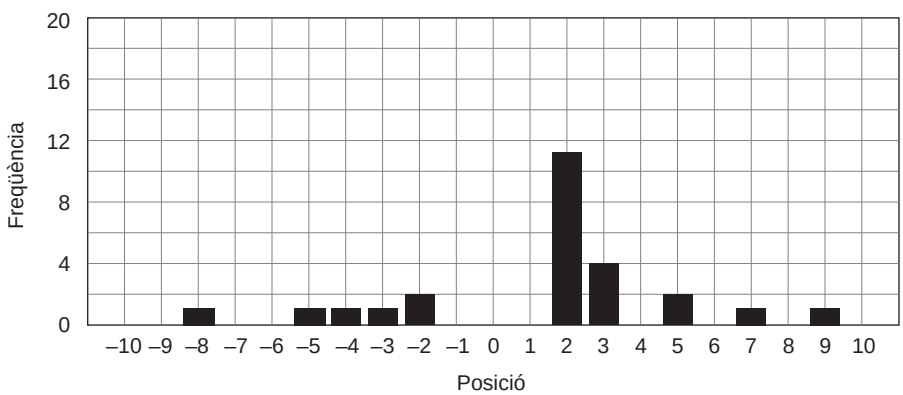


FIGURA 1. Histograma que caracteritza la coocurrència de la forma *platypus* ('ornitorinc', en anglès) i la forma *anatinus* (part de la seva denominació científica). Exemple 1: *…the **platypus** ornithorhynchus **anatinus** is a semiaquatic mammal endemic to eastern Australia, including…*

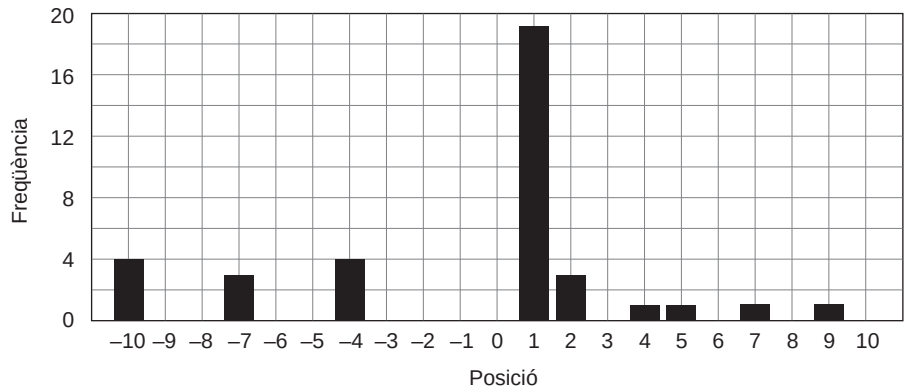


FIGURA 2. Histograma per les formes *platypus* i *has* ('té'). Exemple 2: *…the* ***platypus has*** *four legs which extend horizontally from its body…*

anglesa *platypus* (ornitorinc) en un corpus descarregat d'Internet. En la figura 1, observem que les ocurrències de la forma *anatinus,* una de les paraules amb les quals està associada, es reparteixen a esquerra i dreta de *platypus*. Comprovem aquí que les ocurrències de *anatinus* es concentren en la posició +2, és a dir que la majoria de les vegades la forma *anatinus* apareix dues posicions després de la forma *platypus*, com en l'exemple 1. En la figura 2, observem que el mateix passa amb la forma *has*, tot i que ara la forma es concentra en la posició +1, tal com ocorre en l'exemple [2].

En lingüística de corpus és habitual utilitzar mesures d'associació, però no tant per a falsar una hipòtesi nul·la, segons la qual els elements que estem estudiant es combinen per atzar, sinó més aviat per a ordenar combinacions d'elements a partir de la ponderació que obtenen a conseqüència de l'aplicació d'aquestes mesures. Podem establir diferents tipus de mesures d'associació en funció de la simetria o asimetria que presenten. Entre les mesures d'associació simètriques trobem el concepte d'informació mútua (fórmula [6]), derivat de la teoria de la informació. Representa la quantitat d'informació que ens dóna l'ocurrència de l'esdeveniment *i* sobre l'ocurrència de l'esdeveniment *j* (Church i Hanks, 1991; Manning i Schütze, 1999). Amb aquesta fórmula mesurem, en bits, com és de previsible un esdeveniment *i* en passar *j*, és a dir, quanta sorpresa ens causa *i* quan apareix *j*. En un cas extrem, una alta informació mútua seria que *i* només passa quan ha passat *j*, i en l'extrem oposat, que si passa *i* pot passar *j* o qualsevol altre esdeveniment. És simètrica per definició, és a dir, dóna un mateix valor a *i* donada *j*, que a *j* donada *i*. Aquesta mesura no és aplicable a esdeveniments que tenen poca fre-

3. Aquests histogrames es poden generar automàticament amb el programa Jaguar, accessible a través d'Internet (*http://jaguar.iula.upf.edu*).

qüència, ja que atorgaria una alta associació als que apareixen en conjunció per simple atzar.

$$\mathrm{MI}(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)} \qquad [6]$$

Entre les mesures d'associació asimètriques trobem la probabilitat condicional dels esdeveniments *i* i *j* (fórmula [7]). És una mesura asimètrica, perquè pot no ser igual la probabilitat *i* donada *j*, que la probabilitat de *j* donada *i*. Per exemple, si *j* és la paraula *auguri* i *i* és *mal* (o *bon*), la paraula *auguri* prediu *mal*, però *mal* no prediu en absolut *auguri*.

$$p(i \mid j) = p(i \cap j) \,/\, p(j) \qquad [7]$$

Fins ara hem vist exemples amb bigrames, és a dir, seqüències de dues paraules. Si estem estimant la probabilitat d'aparició d'un bigrama, podríem també tornar a la fórmula [1] i definir-ne la probabilitat com la freqüència d'aparició dividida per la quantitat total de bigrames que hem observat en un corpus.

Veurem en la secció 4 que és possible, estudiant només les freqüències d'aparició dels bigrames, reconèixer l'escriptura d'autors individuals. Això és posible, perquè el llenguatge és un sistema d'opcions i eleccions. El llenguatge ofereix al parlant o a l'autor diferents possibilitats de combinatòria, i aquest últim, amb les seves eleccions, es va construint a si mateix. Llavors hi comença a haver combinacions que són recurrents o típiques d'un autor en comparació amb altres. Però no parlem només d'autors, perquè també les variants dialectals dels diferents col·lectius o nacions tenen una determinada manera de combinar les paraules i conformen patrons que l'ordinador pot reconèixer mitjançant l'aplicació d'un senzill càlcul estadístic. Aquests patrons, no cal dir-ho, són completament imperceptibles per a l'ull humà.

3.2. *Distribució*

La secció anterior ofereix una visió del corpus com un espai continu on es pot donar la coocurrència d'esdeveniments-paraules, valent-se de la noció de *finestra de context* per a definir quan dues paraules apareixen juntes. Aquesta secció, en canvi, ofereix una perspectiva diferent del corpus, ja que el concebem dividit segons un criteri determinat. En primer lloc, comentarem alguns exemples de com podem estudiar —o, més aviat, visualitzar— la distribució d'unitats o de combinacions d'unitats en corpus dividits de manera diferent. Finalment, estudiarem la manera d'ordenar les unitats d'un corpus a partir del comportament que té la seva corba de distribució.

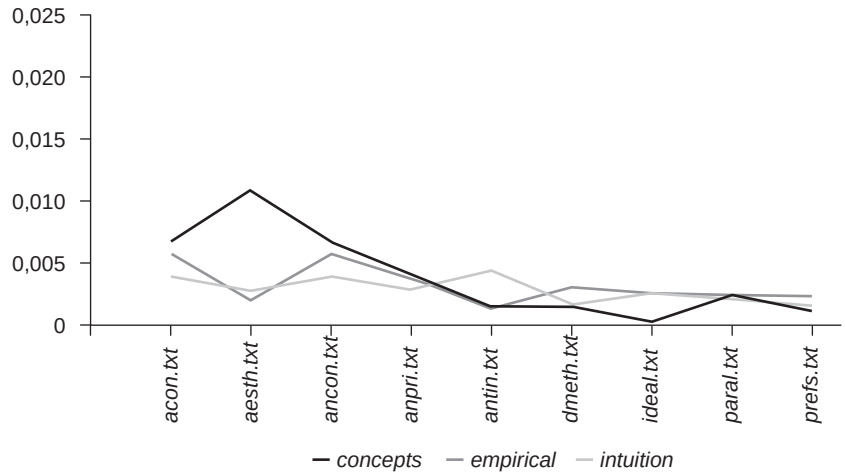


FIGURA 3. Distribució de les formes *concepts*, *empirical* i *intuition* al llarg dels diferents capítols d'una versió en anglès de la *Crítica de la raó pura*, de Kant. L'eix horitzontal representa els diferents capítols. L'eix vertical, la freqüència relativa

El primer exemple és l'anàlisi de la distribució de termes en un document concret. D'acord amb finalitats diverses, ja sigui l'anàlisi del discurs en el pla teòric o l'elaboració de sistemes d'indexació per a la recuperació d'informació, podem tenir interès a esbrinar com es distribueixen les ocurrències de determinats termes en l'obra d'un autor. És possible que existeixin termes clau en certes obres que es distribueixin d'una manera recurrent al llarg del text. També pot ocórrer que alguns termes es concentrin en determinats capítols de l'obra. Pot ser que es trobin en la introducció, per exemple, ja que la seva funció és introduir el lector en els conceptes que després presentarà el text, associats als coneixements que se suposa que té el lector. Però és possible també que aquests termes introductoris no siguin fonamentals en l'obra. La figura 3, per exemple, mostra que tres termes clau en l'obra de Kant, *concepts*, *empirical* i *intuition* es distribueixen de manera regular en l'obra, si bé *intuition* es concentra en el capítol dedicat a l'estètica.

Tanmateix, també és possible que una gran quantitat de paraules es distribueixi de manera regular al llarg de l'obra; però no perquè sigui important per al contingut, sinó perquè forma part del sistema de la llengua. Per això, per als estudis de distribució d'una obra concreta, cal tenir en compte la distribució de les unitats en un corpus. La figura 4 mostra un exemple de distribució d'unitats, aquesta vegada en un corpus diacrònic. Es tracta de les freqüències de les paraules[4] dels ar-

4. Vegeu *http://www.elpais.es*.

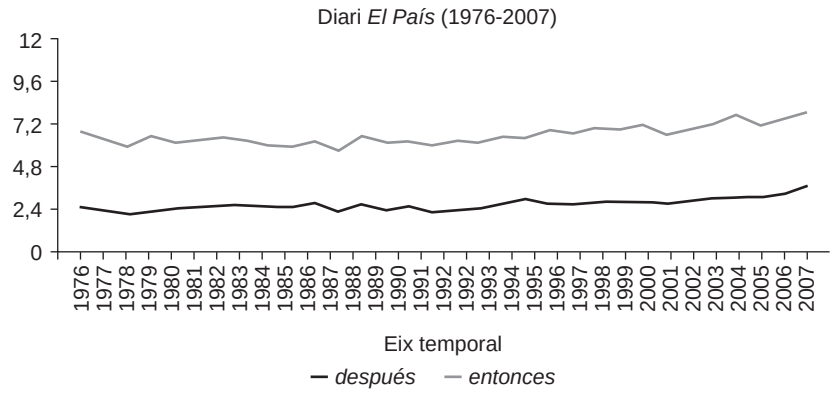


FIGURA 4. Distribució de les formes *después* i *entonces*, dues paraules del vocabulari central de la llengua castellana, en els arxius del diari *El País* en el període 1976-2007. L'eix horitzontal representa el temps i l'eix vertical, la freqüència relativa

xius del diari *El País*.[5] Cadascuna de les divisions en l'eix horitzontal representa totes les edicions d'un mateix any. L'eix vertical representa la freqüència relativa d'una paraula determinada o d'una combinació de paraules en cada any. Podem observar que, mentre que algunes paraules tenen un ús continu al llarg del temps, ja que són paraules del vocabulari central de la llengua (figura 4), altres unitats tenen un ús que fluctua, ja que fan referència a conceptes extralingüístics que tenen diferent vigència en funció de l'agenda temàtica dels mitjans de comunicació (figura 5).

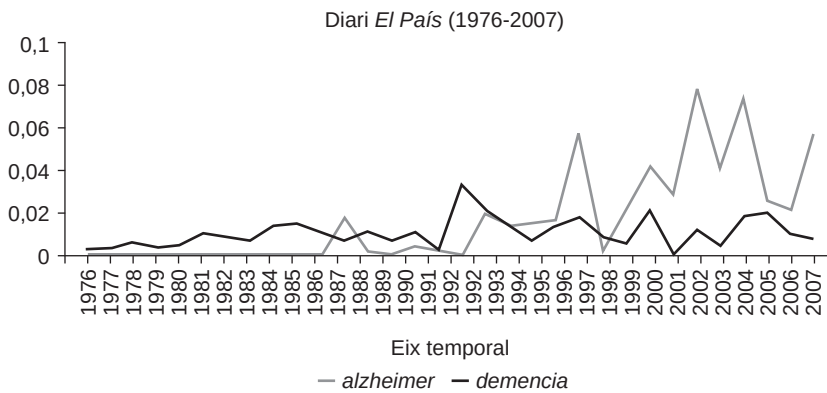


FIGURA 5. Distribució de les formes *demencia* i *Alzheimer* en el mateix corpus, dues unitats que fan referència a coneixement extralingüístic

5. El programa que genera aquestes gràfiques es pot consultar per Internet a l'adreça *http://melot.upf.edu/elpais*.

Un cas diferent és el de dues unitats que, si bé també estan implantades, presenten oscil·lacions a causa de l'evolució del sistema semàntic de la llengua. Ho exemplifiquem en la figura 6, amb les unitats *hombre* i *mujer*, que representen el desenvolupament ideològic d'una societat que pren consciencia del llenguatge sexista. Així, veiem que mentre l'any 1976 la paraula *hombre* és molt més comuna que la paraula *mujer*; aquesta diferència es va revertint amb el temps fins a assolir la mateixa freqüència d'ús l'any 2007.

Basant-nos en el comportament de les corbes de distribució de freqüències de les unitats en aquests corpus dividits, hi ha diversos coeficients que ens interessen per diferents finalitats. En alguns casos, ens interessaran les unitats o combinacions d'unitats que tinguin una freqüència d'ús ascendent, com en el cas de l'extracció de neologia (subsecció 4.3). Però en altres casos ens interessarà saber quin és el vocabulari consolidat d'una llengua, per contrast amb les unitats referencials, és a dir, aquelles que fan referència a coneixement extralingüístic. En aquest cas, ens interessen aquelles unitats que tinguin les corbes més horitzontals. En el cas oposat, podem caracteritzar la irregularitat d'una distribució mitjançant la fórmula [8] (Nazar, 2008) que mesura la dispersió D d'una unitat t per mitjà de la multiplicació del valor màxim de freqüència de t o $\max f(t)$, que seria la freqüència de t en la partició on és més freqüent, multiplicada per $\mathrm{Cr}(t)$, que seria la quantitat de particions en què t té freqüència 0 o una freqüència inferior a un paràmetre k.

$$D(t) = \max f(t) \cdot \mathrm{Cr}(t) \qquad [8]$$

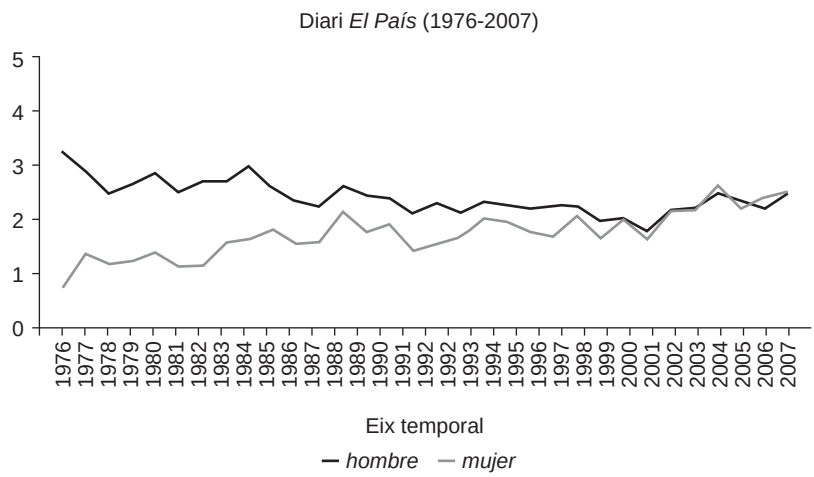


FIGURA 6. Distribució de les formes *hombre* i *mujer* en el mateix corpus

3.3. *Similitud*

En aquesta secció tractem el concepte de *similitud* des d'un espectre ampli. Podríem parlar exclusivament de *similitud entre entitats lingüístiques*, però cal saber que també és possible calcular la similitud entre diferents objectes complexos si som capaços de codificar-los com a vectors. Podem agrupar diferents objectes segons la similitud que tinguin, definida d'acord amb els atributs que comparteixin. Aquests atributs estaran definits per a cada objecte en forma de vector. Un vector pot representar diverses coses: un document, el feix de coocurrències d'un terme, els predicats amb els quals acostuma a aparèixer un nom, etc. La quantitat de valors d'un vector és el que determina la seva dimensionalitat, n, on els x_i en són els components (fórmula [9]).

$$\vec{x} = (x_1, x_2, x_3, \ldots, x_n) \qquad [9]$$

Un vector s'intueix amb facilitat com una fila d'una matriu. La taula 1 mostra, per exemple, una matriu de document per terme, mentre que la taula 2 mostra una matriu de terme per terme.

TAULA 1. *Matriu de document per terme*

	$Term_1$	$Term_2$	$Term_3$	...
Doc_1	1	0	1	...
Doc_2	0	1	1	...
Doc_3	0	1	0	...
...	...	...	...	...

TAULA 2. *Matriu de terme per terme*

	$Term_2$	$Term_3$	$Term_4$	...
$Term_1$	1	0	1	...
$Term_2$	–	0	1	...
$Term_3$	–	–	0	...
...	...	...	...	...

Si els objectes que estem comparant fossin termes i els components dels seus vectors representessin els n-grames de lletres que els conformen, llavors podríem

utilitzar les mesures de similitud entre cadenes de caràcters pel fet de tenir, entre altres coses, una forma de pseudolematització en el treball amb textos no etiquetats, ja que aquesta metodologia seria capaç de detectar la similitud que existeix entre cadenes com *malaltia* i *malalties*; o bé la identificació de variants terminològiques, com en el cas de *superfície pulmonar* i *superfície dels pulmons*.

Amb mesures de similitud com aquestes podem elaborar, per exemple, un programa que, a partir d'un terme d'entrada, indiqui una llista de termes en un corpus que presenten una similitud morfològica. El mateix es pot fer amb documents: a partir d'un document determinat, el programa ordenarà la resta dels documents del corpus d'acord amb la similitud. Però les possibilitats no es limiten a això. En la seva tesi, per exemple, Vanesa Vidal (en preparació) té un experiment en el qual compara diferents verbs especialitzats en funció dels noms amb els quals aquests verbs solen aparèixer.

TAULA 3. *Matriu de verbs per noms*

	Nom_1	$Norm_2$	Nom_3	Nom_4	...
$Verb_1$	0	0	0	1	...
$Verb_2$	1	0	0	0	...
$Verb_3$	0	0	1	0	...
...	...	...	...	...	...

La taula 3 mostra un fragment d'una matriu que té centenars de files i columnes que encreuen la informació de coocurrència de verbs (files) i noms (columnes) en un corpus de genoma. És una matriu binària, ja que codifica, en cada cel·la, l'aparició o la no-aparició de les combinacions verbonominals. La comparació automàtica de tots els verbs[6] entre si dóna una llista dels grups de verbs més similars, és a dir, aquells que es relacionen amb el mateix o gairebé amb el mateix grup de noms. D'aquesta manera, podrem veure que, sense tenir en compte cap tipus d'informació sobre la similitud morfològica i ortogràfica, trobem que, en castellà, en l'àmbit de genoma, els verbs *enrollar* i *desenrollar* són molt semblants perquè apareixen al costat dels noms *hélice*, *cadena*, *adn*, *hebra*, etc.; així com els verbs *beber, ingerir* i *reabsorber* s'assemblen perquè comparteixen els noms *agua*, *cantidad*, *cola, célula* i *glucosa*, entre d'altres. Diferents autors han adoptat estratègies més o

6. El programa que fa aquesta comparació (algorisme de *clustering*) es pot executar a través d'Internet a l'adreça *http://melot.upf.edu/clusteau*, però encara no està suficientment documentat.

menys semblants; no ja en l'estudi de combinacions verbonominals, sinó per al descobriment de sinònims, quasisinònims o bé equivalents en diferents llengües que posen en relació elements que comparteixen els mateixos veïns (Nazar, en preparació).

Entre altres mesures de similitud, la mesura Dice és apropiada per a la comparació de vectors amb valors binaris. El que fa és comptar la quantitat de dimensions en què en dos vectors el valor és superior a zero. Si X i Y són els dos vectors, la mesura queda expressada en la fórmula [10]. $|X|$ és el conjunt cardinal de X, és a dir, la quantitat de components. Es multiplica per dos per tenir un escala que va de 0,0 a 1,0, que seria la similitud total.

$$\text{Dice}(X, Y) = \frac{2\left|X \cap Y\right|}{\left|X\right| + \left|Y\right|} \qquad [10]$$

La mesura Jaccard (fórmula [11]) és similar a l'anterior, però introdueix una normalització: la divisió per la quantitat de dimensions dels vectors, és a dir, que introdueix una penalització quan hi ha poques dimensions compartides en proporció a la quantitat total de dimensions.

$$\text{Jaccard}(X, Y) = \frac{\left|X \cap Y\right|}{\left|X \cup Y\right|} \qquad [10]$$

4. Aplicacions pràctiques

Si bé la secció anterior ja suggereix alguns exemples d'aplicació pràctica, en aquesta secció presentem un espectre d'aplicació més ampli. Analitzarem l'aplicació de mesures de similitud i coocurrència en l'àmbit de la classificació automàtica de documents en les dues modalitats en què aquesta pràctica existeix actualment: la classificació amb aprenentatge supervisat i no supervisat. Finalment, comentarem breument l'aplicació de mesures de distribució aplicades al descobriment de neologia. La manca d'espai ens obligarà a deixar temes que hauria estat molt interessant comentar, com per exemple l'aplicació de metodologies estadístiques a l'extracció de terminologia especialitzada, així com les metodologies per a l'extracció de terminologia bilingüe de corpus no paral·lels, que són línies de recerca en curs.

4.1. *Classificació de documents*

Com és sabut, els algorismes de classificació automàtica de documents es divideixen en *supervisats* i *no supervisats* (Manning i Schütze, 1999; Sebastiani, 2002). En ambdós casos estem agrupant objectes (documents, en aquest context), però la diferència és que, en el primer, un algorisme de classificació té un coneixement previ sobre els objectes que ha de classificar, ja que ha passat per un procés d'«entrenament», en el qual un usuari li ha ensenyat exemples d'objectes classificats segons un criteri qualsevol. En el segon cas, en canvi, la tasca de classificació s'ha de fer sense aquest coneixement, és a dir que l'algorisme no sabrà quantes ni quines són les categories segons les quals els objectes han de ser agrupats, i per tant la classificació serà una propietat que sorgirà a partir de les similituds que tenen els objectes.

4.1.1. Classificació amb aprenentatge supervisat

L'any 2004 em vaig vincular a dos grups d'investigació que estaven treballant en àrees que en principi poden semblar dissímils. Un dels grups estava treballant en l'atribució d'autoria amb el propòsit d'aplicar-la a la lingüística forense. L'altre grup, més vinculat a la terminologia, tenia interès a trobar una manera sistemàtica de classificar un document, tant segons la temàtica com segons el grau d'especialitat. La filosofia de treball en ambdós grups era la mateixa: dissenyar estratègies fonamentades en el coneixement lingüístic, entesa com l'examen manual de la casuística i la identificació, d'acord amb la intuïció de l'investigador, d'aquells trets que podrien ser discriminants de les diferents categories. En ambdós casos es tracta d'un treball d'enorme complexitat i arrelat en el coneixement que l'investigador té de la llengua particular en què està escrit el text. En el cas de la lingüística forense, aquests trets poden ser, per esmentar alguns exemples, girs idiosincràtics que puguin delatar una pertinença a una zona geogràfica o a una condició social, o bé particularitats com els errors d'ortografia o gramàtica que tinguin en comú els textos d'autoria disputada amb aquells textos d'autoria indubtable (vegeu Turell, 2005, per a una introducció). En el cas de la classificació de documents per tema o per grau d'especialitat, l'estratègia consistia a trobar trets lingüístics d'un domini temàtic (la densitat de terminologia especialitzada en el text, per exemple) o bé altres trets morfològics i lèxics que poden ser característics de la literatura especialitzada (Cabré *et al.*, 2009).

En aquest context, va sorgir el programari Poppins.[7] Aquest programa representa una solució de classificació diferent, ja que es pot aplicar tant als problemes d'atribució d'autoria com a la classificació per tema, per grau d'especialitat i fins i

7. El programa Poppins pot ser executat a través d'Internet a l'adreça *http://www.poppinsweb.com*.

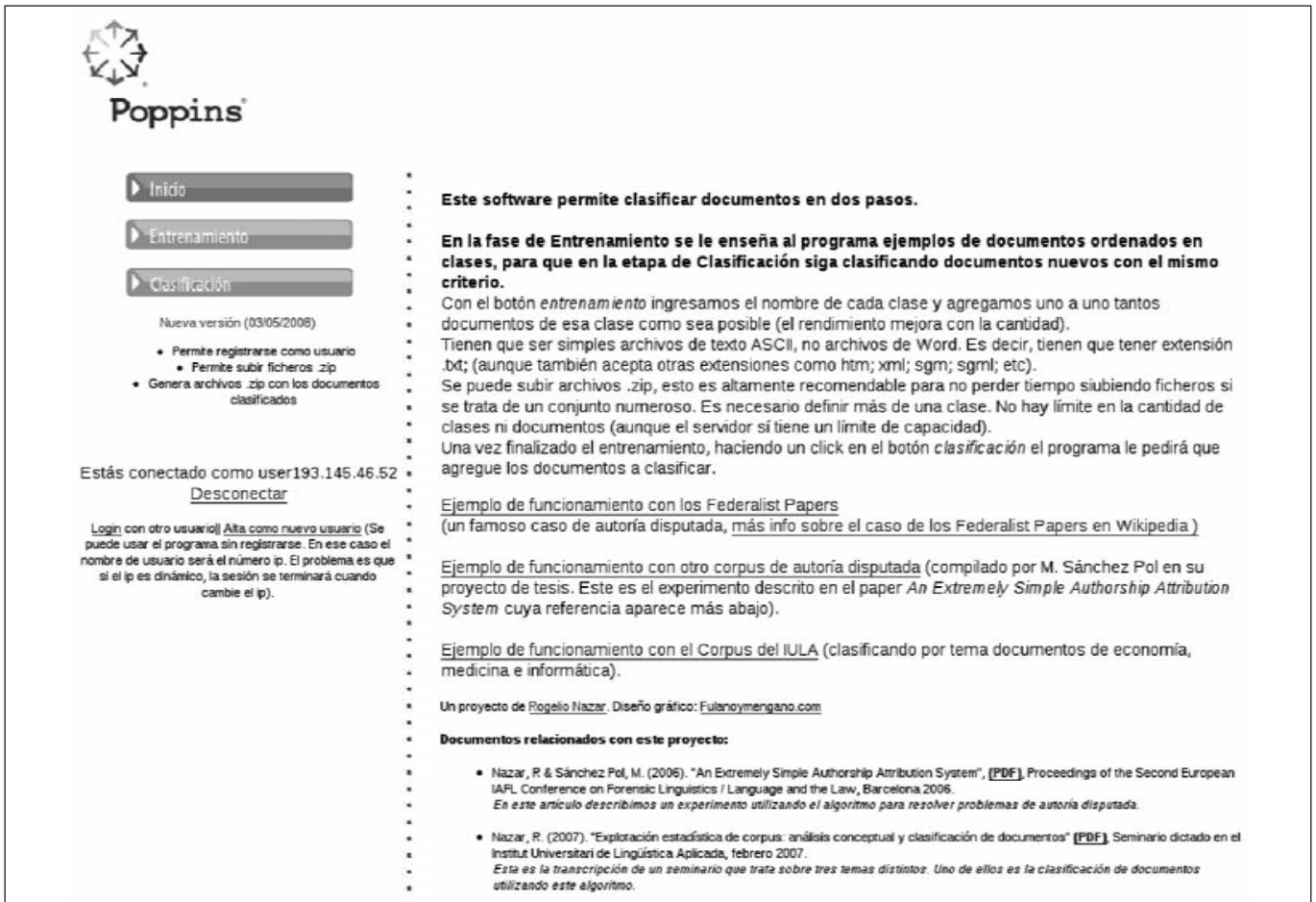


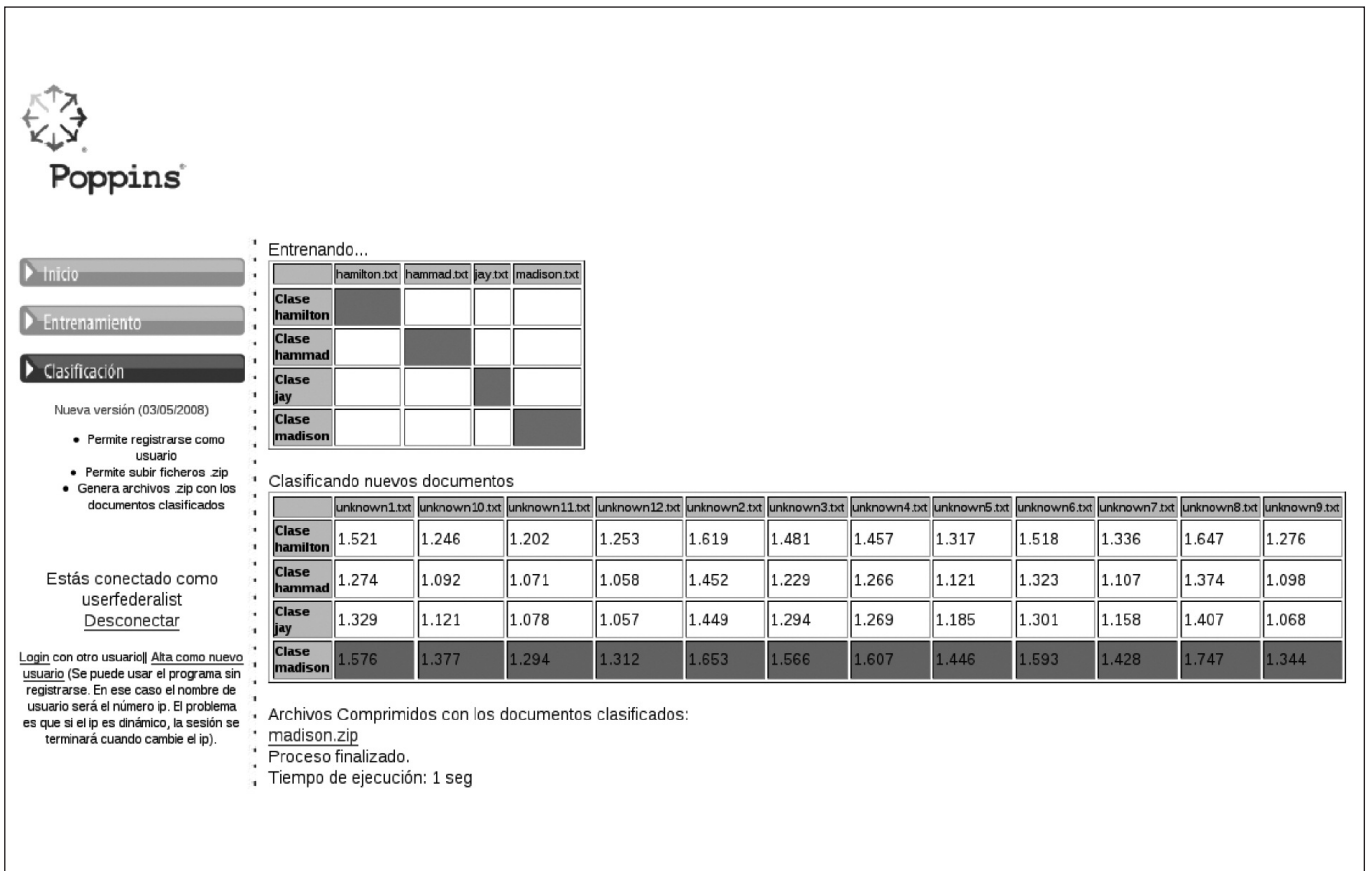
FIGURA 7. Interfície web del programa de classificació automàtica Poppins (*http://www.poppinsweb.com*)

tot per altres problemes de classificació en els quals l'algorisme sigui entrenat, i això amb independència de la llengua dels documents, del domini temàtic o del criteri de classificació. Com dèiem abans per al cas dels algorismes supervisats, la lògica d'aquest programa inclou dues fases principals. A la primera, la fase d'entrenament, un usuari «presenta» al programa exemples de documents ordenats en classes. Un cop acabada aquesta etapa, l'etapa de classificació consisteix a, partint d'un nou conjunt de documents, ordenar-los basant-se en la classificació que ha après durant la fase d'entrenament. El mode de funcionament és bàsic perquè els textos que són classificats no són sotmesos a cap tipus de processament. L'única operació que es fa és calcular les freqüències d'aparició dels diferents bigrames de paraules del corpus. Així, cada classe d'entrenament es converteix en un vector que té per atributs els bigrames, i, per valor, la freqüència d'aparició. D'aquesta manera, a partir d'un nou document, el que fem és computar una mesura de similitud que consisteix a sumar les freqüències dels bigrames que tenen en comú el document per classificar i cadascuna de les classes. La comparació que obté com a resultat la suma més gran és la classe escollida per a aquest document.

Amb Marta Sánchez Pol (Nazar i Sánchez Pol, 2006) vam descobrir que, amb aquest programa, podíem determinar correctament l'autoria d'un text amb una probabilitat del 90 %. La interfície del programa mostra experiments amb altres casos, com el dels Federalist Papers, un famós cas d'autoria disputada, i atribueix els textos d'autoria disputada a James Madison (figura 8) tal com han demostrat altres estudis duts a terme (Mosteller i Wallace, 1984). Pel que fa a la classificació per temàtica i per grau d'especialitat, experiments de classificació de documents del Corpus Tècnic de l'IULA van demostrar nivells de precisió semblants. L'experiment encara es pot repetir de diverses maneres, mitjançant la classificació dels documents per llengua, per variant dialectal o per altres criteris.

4.1.2. Classificació amb aprenentatge no supervisat

Com hem dit en la introducció d'aquesta secció, la classificació amb aprenentatge no supervisat és l'escenari en el qual l'algorisme no ha passat per una etapa d'entrenament i, per tant, no sap quines ni quantes són les categories en què han de ser classificats els documents. Si en el cas anterior relacionàvem la classificació de documents amb aplicacions concretes com l'atribució d'autoria, en aquest cas la classificació de documents amb aprenentatge no supervisat es relaciona amb la desambiguació de terminologia. Això és així perquè plantegem la classificació com un problema de desambiguació. En aquest experiment, reunim una col·lecció de documents en què apareix una forma ambigua, per exemple per mitjà de la descàrrega de documents d'Internet, i els classifiquem a partir dels diferents sentits que pot mostrar aquesta forma dins la col·lecció. Aquesta classifica-



	hamilton.txt	hammad.txt	jay.txt	madison.txt
Clase hamilton				
Clase hammad				
Clase jay				
Clase madison				

	unknown1.txt	unknown10.txt	unknown11.txt	unknown12.txt	unknown2.txt	unknown3.txt	unknown4.txt	unknown5.txt	unknown6.txt	unknown7.txt	unknown8.txt	unknown9.txt
Clase hamilton	1.521	1.246	1.202	1.253	1.619	1.481	1.457	1.317	1.518	1.336	1.647	1.276
Clase hammad	1.274	1.092	1.071	1.058	1.452	1.229	1.266	1.121	1.323	1.107	1.374	1.098
Clase jay	1.329	1.121	1.078	1.057	1.449	1.294	1.269	1.185	1.301	1.158	1.407	1.068
Clase madison	1.576	1.377	1.294	1.312	1.653	1.566	1.607	1.446	1.593	1.428	1.747	1.344

FIGURA 8. Resultat de la classificació de documents per autor mitjançant el programa Poppins en el cas de l'autoria disputada dels Federalist Papers

ció es duu a terme per mitjà dels grafs de coocurrència lèxica. Prenguem, en primer lloc, un exemple amb una forma ambigua com *ratón*, en castellà, que, en el Corpus Tècnic de l'IULA (Vivaldi, 2009) —que conté documents sobre informàtica i sobre genoma— pot ser utilitzada per a fer referència a un dispositiu perifèric de l'ordinador o bé a un animal de laboratori.

En els grafs de coocurrència hi ha un node principal, situat a la zona superior central, que és el terme que estem analitzant: *ratón*, en aquest cas. D'aquest node, en depenen tots els altres. Cada node representa una paraula o una combinació de paraules, i les connexions entre nodes expressen que les paraules que els nodes representen apareixen juntes en els mateixos contextos on apareix la unitat que estem analitzant. En la figura 9 s'aprecia l'existència de dues regions en el graf, una a la dreta i una altra a l'esquerra. Aquestes dues regions —atractors o clústers de nodes— es corresponen amb cadascun dels sentits que la forma presenta en el corpus. En un cas, les unitats amb les que apareixerà *ratón* seran *cromosoma, mamífero, rata, genoma, laboratorio, bacteria,* entre altres; mentre que, en l'altre cas, les unitats que es relacionen amb *ratón* són *usuario, pantalla, teclado, clic,* etcètera.

En la tesi (en preparació) presento, entre altres coses, un estudi de desambiguació de sigles, ja que aquestes són formes ambigües per naturalesa. Així, davant d'una col·lecció de documents descarregada d'Internet amb una forma ambigua com *NLP*, per exemple, un programa informàtic[8] és capaç d'obtenir dos clústers que representen els dos sentits d'aquesta paraula: d'una banda, documents referits a la forma expandida *natural language processing* i de l'altra, documents sobre *neuro-linguistic programming*. En el primer cas, *NLP* es relaciona amb unitats com *knowledge representation*, *language technology*, *functional grammar*, *machine translation*, *statistical NLP*, *computational linguistics*, entre altres; mentre que el segon clúster inclou unitats com *practitioner training*, *practitioner NLP*, *gestalt therapy*, *John Grinder*, *Richard Bandler*, *Robert Dilts*, etcètera.

4.2. *Descobriment de neologia*

En aquesta secció analitzarem l'aplicació d'algunes de les mesures de distribució que hem vist en la secció 3.2, amb el propòsit concret de fer un experiment d'extracció automàtica de neologia. Els resultats de l'aplicació d'aquestes tècniques per a l'extracció de neologia, així com de les tècniques de desambiguació automàtica presentada en el punt anterior (4.1.2) van ser presentades en un treball previ (Nazar i Vidal, 2008).

8. El programa que fa aquesta classificació de documents descarregats d'Internet mitjançant una forma ambigua també es pot executar a l'adreça *http://melot.upf.edu/mandinga*, encara que no existeix documentació per al programa i la interfície és encara rudimentària. El resultat de l'experiment de *NLP* es pot veure a la adreça següent: *http://melot.upf.edu/nlp*.

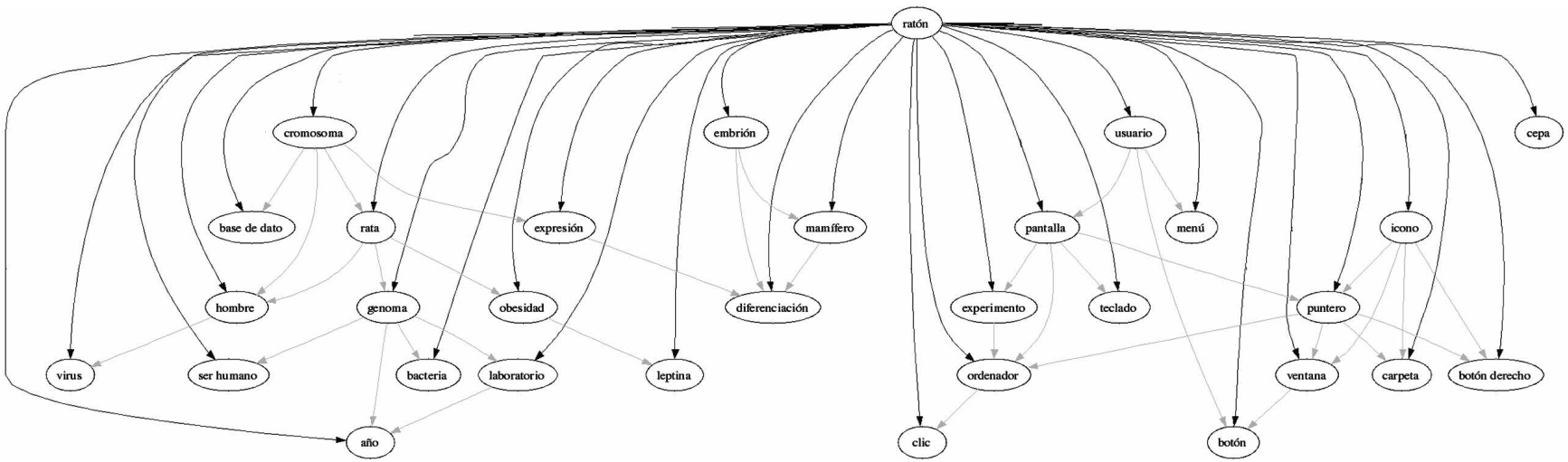


FIGURA 9. Graf de coocurrència de la forma ambigua *ratón*, utilitzada per a fer referència al dispositiu perifèric de l'ordinador en els documents d'informàtica i a l'animal utilitzat en laboratori en els documents de genoma

Les figures 10 i 11 ofereixen gràfiques que ja ens són familiars, perquè hem vist corbes semblants en la subsecció 3.2: seguiments de determinades unitats lèxiques al llarg del corpus diacrònic d'*El País*. Mostren exemples del comportament d'unitats que considerem neologismes, com ara *teléfono móvil*, *teléfono fijo* i *cambio climático*, unitats la freqüència d'ús de les quals mostra un increment acusat en la línia del temps.

$$f(x) = x^{10} \qquad [12]$$

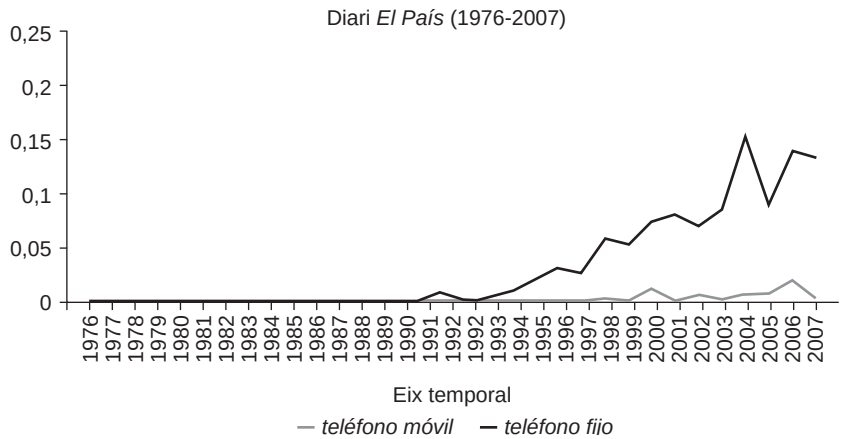


FIGURA 10. Distribució de les formes *teléfono móvil* (corba superior) i *teléfono fijo* (corba inferior) en el corpus diacrònic d'*El País*

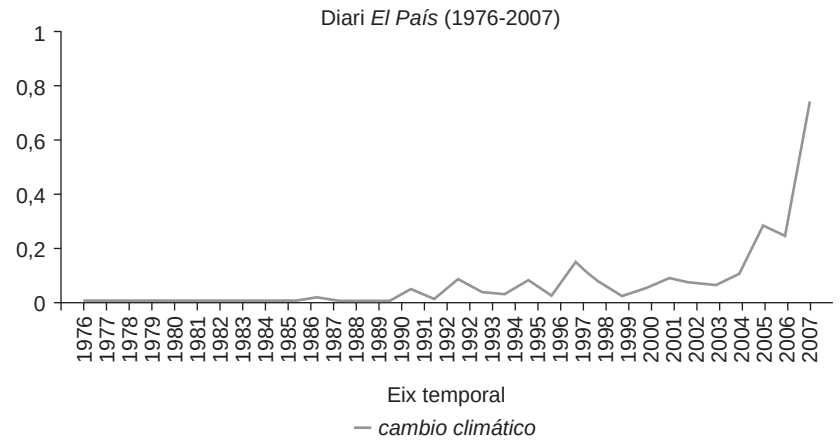


FIGURA 11. Distribució de la forma *cambio climático* en el mateix corpus

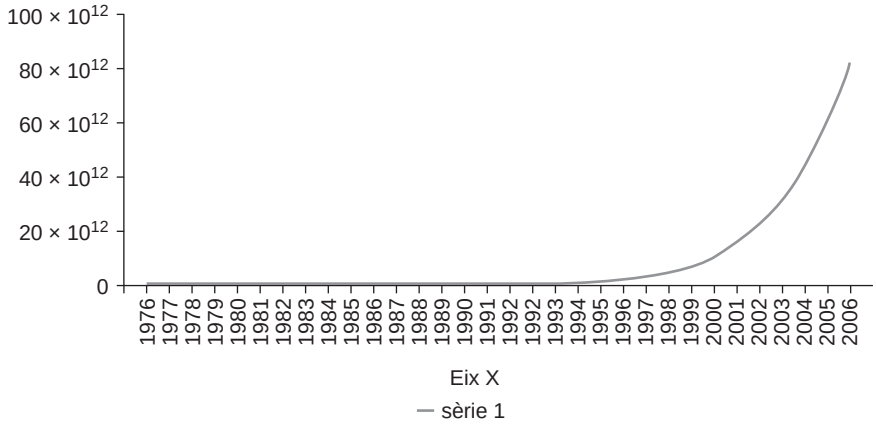


FIGURA 12. Gràfica del neologisme ideal

En l'esmentat treball sobre extracció de neologia vam definir el que seria la corba de comportament d'un neologisme ideal o teòric, representada en la figura 12 i definida en la fórmula [12]. Es tracta d'una corba exponencial en l'interval d'anys estudiat. L'experiment va consistir a prendre una mostra de *n* unitats del corpus (les unitats eren tant paraules aïllades com seqüències de fins a cinc paraules de longitud) i a ordenar-les d'acord amb la distància euclidiana de les seves corbes de freqüència amb la corba d'aquest neologisme ideal. D'aquesta manera, vam obtenir les unitats que s'han anat incorporant a la llengua en els darrers anys, unitats que després s'han de filtrar, ja que inclouen formes que no són neologismes, com és el cas de noms propis o referents que han adquirit notorietat en els darrers anys.

Naturalment, aquest senzill mètode no resultava eficaç en el cas dels neologismes semàntics, unitats que si bé són formalment idèntiques a altres formes de la llengua, es comencen a fer servir amb un significat diferent. Aquestes formes representen un desafiament per a l'extracció automàtica amb els mètodes tradicionals, però aquest mateix escenari és el que trobàvem en la subsecció 4.1.2, en la qual classificàvem contextos d'aparició d'unitats polisèmiques. És el cas, per exemple, de la forma *palabra de honor*, que si bé té un ús literal, en el sentit de 'fer una promesa verbal', en els darrers anys és cada vegada més freqüent utilitzar-la per a designar un determinat tipus d'escot. Si bé la seva condició de neologisme per a aquest segon sentit és discutible, ja que aquest tipus d'escot no és nou, sí que és nova la massificació d'aquest ús del terme, i, per tant, l'exemple segueix sent útil. Un algorisme de clusterització (*clustering*) similar al descrit en la subsecció 3.2 és capaç de classificar tots els contextos d'aparició de la forma *palabra de honor* en els arxius d'*El País* i oferir dos clústers amb un nom per a cadascun. El clúster 1 és ano-

menat *empeñar* i el clúster 2 és anomenat *escotes*. Cada un d'aquests clústers conté una sèrie d'unitats lèxiques que conformen l'entorn típic de les ocurrències de l'expressió en un sentit i en l'altre. Així, en el clúster 1, tenim unitats com: *Astarloa*, *Barrionuevo*, *confederal*, *consentido*, *credulidad*, *empeñar*, *esclarece*, *Escudero*, *Fusté*, *Herrero*, *incité*, *inocencia*, *proclamar*, *quebrantamiento*, *reiterado*, etc. Aquestes formes es relacionen amb el sentit literal. Veiem que es tracta de noms propis de personatges públics, per als quals la credibilitat no hauria de ser irrellevant. En el cas del segon clúster, en canvi, els veïns típics tenen relació amb el món de la moda: *cubren*, *drapeados*, *escotes*, *Gucci*, *marrón*, *modista*, *ojito*, *organza*, *Swarovski, tonos*, etcètera.

5. Conclusions

Aquest article presenta una visió àmplia de la cruïlla entre la lingüística i l'estadística, i inclou alguns exemples de tècniques que es poden utilitzar per a l'estudi del llenguatge. Aquestes tècniques s'han acompanyat, a més, amb exemples d'aplicació concreta, com és el cas de la classificació de documents amb supervisió o sense, així com la desambiguació de signes polisèmics i el descobriment de neologia. Hauria estat interessant esmentar altres exemples d'aplicació pràctica d'aquestes tècniques, com la utilització de mesures de similitud per a la comparació entre unitats lèxiques de diferents llengües, és a dir, l'extracció de terminologia bilingüe des de corpus no paral·lels, o bé per a la comparació d'unitats lèxiques de diferents varietats dialectals.

A priori, pot semblar que es tracta d'àrees d'aplicació completament diferents, sobretot per a qui està acostumat a enfrontar tasques d'aquest tipus amb la incorporació de regles explícites que codifiquen coneixement de la llengua o del domini temàtic, així com informació semàntica extreta de diccionaris i ontologies, en el cas de l'extracció de terminologia, o corpus d'exclusió lexicogràfics, en el cas de l'extracció de neologia. L'estadística, per contra, possibilita una manera diferent de concebre la llengua. Una investigació de la complexitat, però des d'una perspectiva integradora i simplificadora. Des del punt de vista estadístic, tasques i dades dissímils comencen a semblar relacionades. De vegades, els mateixos mètodes o les mateixes formes de pensar es poden aplicar a problemes que en principi semblaven completament diferents. Concebem, doncs, l'estadística com una «trans-disciplina».

Per a tancar aquest article, és important remarcar que cal no perdre de vista l'aspecte teòric. No estem parlant només de «trucs enginyerils» per resoldre problemes pràctics que no tenen una relació intrínseca amb la lingüística, com si aquestes solucions estiguessin desproveïdes de teoria. Està per veure si l'estadística i la lingüística conformen disciplines diferents o si hi pot haver alguna cosa que anomena-

ríem una «sensibilitat estadística» en l'anàlisi lingüística, una manera d'aproximarnos a les dades, d'advertir patrons, regularitats o tendències en el cúmul dels casos individuals en què l'ull humà no pot veure sinó quantitat i diversitat.

Agraïments

Aquest treball ha estat possible gràcies al finançament per al projecte RICOTERM3 (Ministeri d'Educació i Ciència: HUM2007-65966-C02-01/FILO. Investigadora principal: doctora Mercè Lorente). Voldria agrair també l'ajuda d'Amor Montané i Alba Coll en la redacció en català.

7. Referències

Cabré, M. T.; Bach, C.; Da Cunha, I.; Morales, A.; Vivaldi, J (2009). *Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico*. XXVII Congreso de AESLA (Ciudad Real, 26-28 març 2009).

Castoriadis, C (1975). *La institución imaginaria de la sociedad.* Buenos Aires: Tusquets.

Church, K.; Hanks, P (1990). «Word Association Norms, Mutual Information and Lexicography». *Computational Linguistics*, vol 16, núm. 1, p. 22-29.

Dilthey, W (1986). *Introducción a las Ciencias del Espíritu*. Madrid: Alianza.

Evert, S (2004). *The Statistics of Word Coocurrences*. Tesi (doctorat). Stuttgart: Universitat de Stuttgart. Institut für Maschinelle Sprachverarbeitung, 2004.

Herdan, G (1964). *Quantitative Linguistics*. Washington: Butterworths.

Mandelbrot, B (1961). «On the theory of word frequencies and Markovian models of discourse». A: *Structure of Language and its Mathematical Aspects.* Symposia on Applied Mathematics. American Mathematical Society. Vol. 12, p. 190-219.

Manning, C.; Schütze, H (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Mosteller, F.; Wallace, D (1984). *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. Nova York: Springer.

Muller, C. (1973). *Estadística Lingüística*. Madrid: Gredos.

Nazar, R. (2008). Diferencias cuantitativas entre referencia y sentido. Actas del XXVI Congreso de AESLA. (Universitat d'Almeria, 3-5 d'abril de 2008).

— ([en preparació]). *Quantitative Approach to Concept Analysis*. Tesi (doctorat). Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

Nazar, R; Sánchez Pol, M. (2006). *An Extremely Simple Authorship Attribution System*. Second European IAFL Conference on Forensic Linguistics / Language and the Law (Barcelona, 2006).

Nazar, R.; Vidal, V. (2008). *Aproximación cuantitativa a la neología*. I Congreso Internacional de Neología en las lenguas románicas (Barcelona, 7-10 maig 2008).

SEBASTIANI, F. (2000). *Machine earning in automated text categorization*. ACM Press, vol. 34, núm. 1.

SHANNON, C. E. (1948). «A mathematical theory of communication». *Bell System Technical Journal*, vol. 27 (juliol), p. 379-423.

SNOW, C. P. (1959 [1993]). *The Two Cultures*. Cambridge: Cambridge University Press.

TURELL, M. (2005). «Presentación». A: *Lingüística forense, lengua y derecho: conceptos, métodos y aplicaciones.* Barcelona: Universitat Pompeu Fabra. Institut Universitari de Ligüística Aplicada, p. 13-16.

VICKERS, B. (2002). *Counterfeiting Shakespeare*. Cambridge: Cambridge University Press.

VIDAL, V. (en preparació) *Combinatoria verbo-nominal en el discurso de especialidad. Delimitación, caracterización y soluciones terminográficas.* Tesi (doctorat). Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

VIVALDI, J. (2009). *Corpus and exploitation tool*: *IULACT and bwanaNet*. I Congreso Internacional de Lingüística de Corpus (Múrcia, 7-9 maig 2009).

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 75-83 DOI: 10.2436/15.2503.02.3

Ús d'estratègies estadístiques per a l'extracció automàtica d'unitats terminològiques

Mercè Vázquez, Antoni Oliver
Universitat Oberta de Catalunya
Barcelona

Resum

La detecció automàtica d'unitats lèxiques de caràcter especialitzat d'un determinat àmbit de coneixement és un dels reptes clau per a l'organització i la recuperació d'informació. En aquesta comunicació es planteja l'ús de diferents estratègies estadístiques, amb l'objectiu de poder extreure automàticament unitats terminològiques d'un àmbit d'especialitat a fi de recuperar i organitzar la informació que conté.

Paraules clau: classificació de documents, documentació, extracció de terminologia, mètodes estadístics, ontologia, recuperació d'informació, taxonomia.

Abstract: ***The use of statistics-based strategies for the automatic extraction of terminology units***

The automatic detection of lexical units of a specialised nature in a given area of knowledge is one of the key challenges in the organisation and retrieval of information. This communication addresses the use of different statistics strategies with a view to be able to automatically extract terminological units from a specialist area to retrieve and organise the information it contains.

Key words: classification of documents, documentation, terminology extraction, statistical methods, ontology, information retrieval, taxonomy.

1. Terminologia i documentació

Els àmbits de coneixement de la terminologia i la documentació s'han centrat, d'una banda, en la identificació i compilació dels termes i, de l'altra, en la identificació i compilació dels documents. Aquestes dues branques de coneixement han anat avançant en la descoberta de noves tècniques per a millorar llurs

processos de treball, però han tingut poques ocasions de compartir l'expertesa assolida en cada una de les àrees. En els darrers anys s'ha vist la necessitat de començar a compartir coneixement i establir lligams entre els especialistes d'aquestes dues àrees per a assolir resultats que puguin ser aprofitats en les dues àrees de coneixement.

En l'àmbit de la terminologia, el reconeixement automàtic d'unitats terminològiques i la detecció precoç de neologismes són alguns dels reptes que encara té pendents actualment el treball terminològic, els quals constitueixen la base de la proposta que es fa en el present article. La tasca de detecció automàtica d'unitats terminològiques i la compilació d'aquestes unitats permet disposar de material terminològic actualitzat, cada vegada més necessari per l'augment exponencial de recursos digitals, el problema d'accés als continguts i la dificultat que hi ha en l'automatització del contingut dels corpus. En aquest sentit, l'àmbit de la documentació necessita tenir a l'abast recursos terminològics que puguin explotar grans volums de corpus per a poder-ne extreure llistes de paraules clau, útils per a la indexació de continguts; elaborar taxonomies i, en última instància, crear ontologies. Així, doncs, s'estableix un marc d'interacció de coneixement i aprofitament de recursos molt important.

D'altra banda, la introducció d'estratègies estadístiques en el procés d'identificació d'unitats candidates a ser termes fa possible de treballar amb corpus d'especialitat de gran volum que poden ser monolingües, bilingües o multilingües i recuperar els equivalents corresponents de traducció i els contextos d'ús. Els mètodes estadístics reconeixen les unitats terminològiques a partir de la freqüència que tenen en un corpus marcat temàticament. Malgrat ser un càlcul molt senzill, el problema que presenta és que es fa difícil de recuperar termes que apareixen poques vegades en un corpus d'especialitat; per aquest motiu, s'ha de combinar amb l'ús de mesures estadístiques. Així, si es compara el valor de freqüència que té una unitat dins un corpus d'especialitat amb els resultats que ofereixen un conjunt de mesures estadístiques hi ha una evidència superior del caràcter terminològic d'un candidat a terme, ja que mesuren el nivell o grau d'associació de les unitats que constitueixen un candidat a terme.

2. Estratègia d'identificació d'unitats especialitzades

La tria d'una mesura estadística que sigui adequada per a identificar el major nombre de termes d'un corpus d'especialitat segueix un procés de preparació de resultats que comença amb l'extracció automàtica de candidats a terme del corpus i el filtratge d'aquests candidats amb una llista de paraules buides (conjuncions, preposicions, locucions, etc.), a fi de disposar d'una llista de candidats endreçats per freqüència.

clear forward signal 442 903 710 4358 540 671 455
data link layer 322 1589 1554 464 564 322 334
coast earth station 256 279 1007 954 274 256 626
earth station antenna 81 961 1150 279 677 85 97
earth station equipment 51 961 1150 1648 677 57 58
earth station antennas 29 961 1150 149 677 29 37
earth station Hpa 29 961 1150 104 677 29 32
earth station receiver 24 961 1150 134 677 24 35
earth station transmit 21 961 1150 81 677 24 21
earth station identification 16 961 1150 291 677 16 33
earth station HPàs 14 961 1150 35 677 16 16
earth station receive 13 961 1150 84 677 13 13
earth station complexes 12 961 1150 12 677 12 12
earth station located 10 961 1150 134 677 19 10
earth station transmitter 10 961 1150 42 677 10 10
earth station owner 8 961 1150 8 677 8 8
earth station number 2 961 1150 994 677 2 42

FIGURA 1. Llista de candidats a terme ordenats per valor de freqüència

En la imatge superior (figura 1) observem una mostra de candidats a terme filtrats amb una llista de paraules buides i endreçats per freqüència que pertanyen a un corpus d'especialitat de l'àmbit de les telecomunicacions. El valor de freqüència correspon al primer valor que apareix al costat del candidat a terme. La resta de valors corresponen al nombre de vegades que apareixen juntes en el corpus les diferents paraules que formen el candidat a terme. Així, el candidat «clear forward signal» veiem que apareix 442 vegades en el corpus, i així successivament.

A partir d'aquí, aquest resultat és processat amb tretze mesures estadístiques[1] que permeten de calcular una puntuació i un valor de rang per a cada candidat i mostren el resultat obtingut en ordre ascendent. La puntuació que s'atribueix a cada candidat indica si hi ha evidència o no n'hi ha que pugui ser una unitat terminològica.

A partir de la informació de freqüència i la puntuació que s'obté per a cada candidat a terme s'observa en quina posició queda endreçat i també quin valor de rang queda atribuït a cada candidat, tenint en compte que els candidats que tenen una mateixa puntuació queden aglutinats dins un mateix valor de rang. D'aquesta manera, els candidats que tenen un valor de rang més baix i una puntuació més alta corresponen a combinacions poc habituals i, per tant, hi ha una probabilitat més alta que siguin terminològiques. I a la inversa, un valor de rang que sigui alt i

1. Coeficient *Dice*, test *Fishers twotailed*, test exacte de *Fisher left sided*, test exacte de *Fisher right sided*, coeficient *Jaccard*, ràtio *Log-likelihood*, mesura *True mutual information*, mesura *Pointwise mutual information*, ràtio *Odds*, test khi-quadrat de Pearson, mesura *T-score*, mesura *Poisson stirling*, coeficient fi.

que vagi acompanyat d'una puntuació baixa indica que la relació que s'estableix entre les unitats que formen el candidat a terme és més habitual i, en conseqüència, és més probable que es tracti d'una combinació menys específica de l'àmbit d'especialitat o, si més no, més habitual.

clear forward signal 1 0.0389 442 903 710 4358 540 671 455
coast earth station 2 0.0363 256 279 1007 954 274 256 626
data link layer 3 0.0285 322 1589 1554 464 564 322 334
earth station antenna 4 0.028181 961 1150 279 677 85 97
earth station antennas 5 0.0263 29 961 1150 149 677 29 37
earth station Hpa 5 0.0263 29 961 1150 104 677 29 32
earth station receiver 6 0.026224 961 1150 134 677 24 35
earth station transmit 7 0.0260 21 961 1150 81 677 24 21
earth station number 7 0.0260 2 961 1150 994 677 2 42
earth station equipment 8 0.0259 51 961 1150 1648 677 57 58
earth station identification 8 0.0259 16 961 1150 291 677 16 33
earth station HPàs 8 0.0259 14 961 1150 35 677 16 16
earth station complexes 8 0.0259 12 961 1150 12 677 12 12
earth station located 9 0.0258 10 961 1150 134 677 19 10
earth station owner 9 0.0258 8 961 1150 8 677 8 8
earth station receive 10 0.0257 13 961 1150 84 677 13 13
earth station transmitter 10 0.0257 10 961 1150 42 677 10 10

FIGURA 2. Llista de candidats a terme ordenats per valor de rang

En la imatge superior (figura 2) veiem com queda reordenada la llista de candidats a terme després de ser processada per una de les tretze mesures estadístiques esmentades més amunt, concretament els resultats corresponen a la mesura *True mutual information*. Ara el primer valor que hi ha al costat del candidat a terme correspon a la informació de rang, i el següent valor correspon a la puntuació que atribueix aquesta mesura en concret al candidat en qüestió. La informació numèrica restant correspon als valors que hem obtingut en el primer pas del filtratge i que hem comentat en l'exemple anterior, és a dir, a la freqüència i al nombre de vegades que apareixen juntes en el corpus les diferents paraules que formen el candidat a terme. Observem que els candidats que tenen un mateix valor de rang també tenen una mateixa puntuació i queden ordenats consecutivament. Així, el candidat «clear forward signal» ara té un valor de rang 1, una puntuació de 0,0389 i una freqüència d'aparició en el corpus de 442.

Si revisem l'ordre en què han quedat ara reordenats els candidats a terme, veiem que candidats que quedaven situats en les primeres posicions de la llista perquè apareixien amb més freqüència en el corpus ara queden recollits en posicions més baixes que no pas candidats que abans apareixien més avall de la llista de resultats perquè tenien una freqüència més baixa; ens referim concretament als ca-

sos de «data link layer», «earth station equipment» o «earth station receive». Així, doncs, la informació de rang ajuda a precisar el caràcter més o menys terminològic que pot tenir un candidat a terme en una llista de resultats.

El filtratge dels resultats inicials fent ús de tretze mesures estadístiques ha fet possible de comparar els resultats obtinguts i alhora comprovar que hi ha unes mesures que permeten d'endreçar en ordre descendent un nombre més gran d'unitats terminològiques que no pas altres. Així, les mesures que han reordenat un major nombre de termes en les primeres posicions de la llista de resultats són el test *Fisher*, la mesura *T-score* i la mesura *True mutual information*.

3. Tècniques de recuperació d'informació aplicades a l'extracció de terminologia

En l'àmbit de la recuperació d'informació s'apliquen estratègies de localització d'unitats per identificar i classificar continguts que també són molt útils per al procés d'extracció de candidats a terme d'un corpus d'especialitat. En aquest sentit, la mesura que és força utilitzada en recuperació d'informació i que s'ha incorporat a la tasca d'extracció de terminologia és la mesura *tf-idf* (*term frequency - inverse document frequency*), que té per objectiu filtrar els termes que són presents en molts documents. En aquest plantejament, cal quantificar la freqüència d'aparició d'un terme dins un document. Aquest paràmetre, habitualment, es coneix per *factor de freqüència del terme* (*tf*, concepte local) i es considera que dóna una mesura de fins a quin punt aquest terme descriu el contingut del document, és a dir, com més vegades apareix un terme en un document, més pes semàntic té. No obstant això, els termes molt corrents gairebé no aporten la capacitat de distingir si un document és pertinent o no ho és per a una cerca concreta. Per aquest motiu, s'hi introdueix un factor calculat a partir d'una relació inversa respecte a la freqüència d'aparició del terme dins un conjunt de documents (*freqüència inversa de documents*, *idf*), és a dir, la freqüència d'aparició del terme dins un conjunt de documents decreix com més gran és el nombre de documents que en parlen; concepte basat en el corpus. I és que, com més freqüent sigui un terme en el conjunt de documents, menys pes i menys capacitat discriminatòria tindrà i, per tant, representarà, de manera secundària, el conjunt de documents. En canvi, els termes que apareixen poc en el conjunt de documents són els que tindran més pes en la mesura *tf-idf* i, per tant, representaran més bé la totalitat de documents.

En l'àmbit de l'extracció de terminologia, la mesura *tf-idf* és molt productiva per a determinar quins són els termes rellevants d'un corpus d'especialitat. Ara bé, a diferència del que es fa en l'àmbit de recuperació d'informació, la selecció de candidats a terme s'efectua fent servir un corpus de llengua general que serveix per a contrastar les unitats que apareixen en aquest corpus amb les que són

pròpies d'un corpus d'especialitat. En aquest sentit, si un candidat a terme apareix força representat i també força distribuït dins el corpus de llengua general, llavors és descartat com a possible candidat a terme. I, a la inversa, si el candidat no apareix en cap dels àmbits temàtics del corpus de llengua general, hi apareix molt poc o bé queda poc distribuït en els diferents fitxers del corpus, llavors es considera adequat com a candidat a terme. D'aquesta manera, les unitats del corpus d'especialitat que apareixen sovint i força distribuïdes en el corpus de llengua general es considera que corresponen a paraules d'ús general i no pas a paraules pròpies d'un àmbit d'especialitat i, per tant, són descartades com a unitats candidates a ser termes.

En aquest sentit, si reprenem el procés de filtratge que hem comentat més amunt tenint en compte les tècniques de recuperació d'informació aplicades a l'extracció de terminologia, el que fem ara és contrastar la llista de candidats a terme amb el contingut d'un corpus de la llengua general amb l'objectiu de poder obtenir un valor de *tf-idf* per a cada candidat.

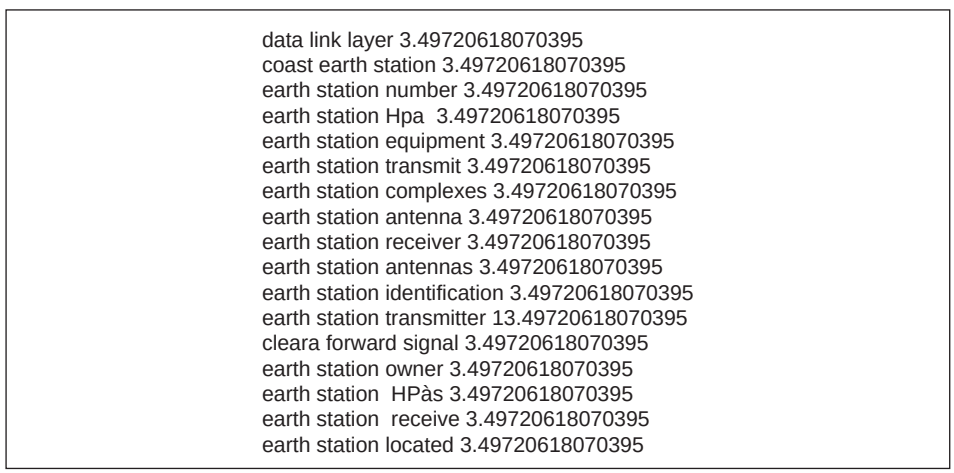


FIGURA 3. Llista de candidats a terme ordenats per valor de *tf-idf*

En la imatge superior (figura 3) podem observar el valor de *tf-idf* que hem obtingut per a la llista de candidats a terme amb què treballem. En aquest cas, el valor de *tf-idf* és igual per a tots els candidats, resultat que ens indica l'alt grau d'especificitat que tenen tots els candidats en ser contrastats amb un corpus de llengua general. També cal tenir en compte que són unitats que apareixen amb molta freqüència en el corpus d'especialitat; per tant, són d'aparició escassa o nul·la en un corpus de llengua general.

4. Combinació d'estratègies en el procés d'identificació d'unitats terminològiques

En el procés d'identificació d'unitats amb caràcter terminològic constatem que la combinació del valor de freqüència d'aparició d'una unitat en un corpus d'especialitat amb els valors de puntuació i rang que ens ofereixen les mesures estadístiques i el valor de *tf-idf*, que és una mesura pròpia de l'àmbit de la recuperació d'informació, permet de classificar millor la llista de candidats a terme tenint en compte el seu caràcter terminològic.

Per aquest motiu, en aquests moments avaluem la possibilitat d'establir un valor de ponderació únic que combini els quatre valors que acabem d'esmentar i, així, poder situar en les primeres posicions dels resultats les unitats que tenen un caràcter terminològic marcat i en les darreres posicions les unitats que són de caràcter menys específic. En aquest sentit, les unitats que tinguin un valor de ponderació més alt seran les que apareixeran amb molta freqüència en el corpus d'especialitat, tindran un valor de rang baix, tindran poca presència en un corpus de llengua general i se situaran en les primeres posicions de la llista de candidats a terme d'un corpus d'especialitat; aquestes unitats tindran un caràcter terminològic marcat i seran susceptibles de formar part d'una llista de termes de referència d'un corpus d'especialitat. I les unitats que tinguin un valor de ponderació més baix correspondran a unitats pròpies d'altres àmbits d'especialitat o bé a combinacions d'àmbit més general que, pel fet de ser usades en un corpus d'especialitat, poden esdevenir unitats específiques de l'àmbit. Així mateix, per poder fer una avaluació objectiva dels resultats que s'obtenen amb un valor de ponderació únic treballem amb una llista de termes de referència propis de l'àmbit d'especialitat del qual s'extreuen els candidats a terme.

En la figura 4 podem observar com queda endreçada finalment la llista de candidats a terme a partir del valor de ponderació. L'ordre en què quedaven endreçats inicialment els candidats amb la mesura *True mutual information* resulta modificat lleugerament després d'haver considerat el valor de *tf-idf* i de freqüència. A tall d'exemple, veiem que el candidat «data link layer», que segons el valor de rang de la mesura estadística *True mutual information* quedava recollit en tercera posició, ara, amb el valor de ponderació únic, queda situat en segona posició, fet que indica que té un major caràcter terminològic que no pas «coast earth station», que ara queda situat en tercera posició. O bé, «earth station Hpa», que amb el valor de ponderació queda situat més amunt en la llista de resultats que no pas amb el valor de rang o amb el valor de freqüència separadament.

clear forward signal 1
data link layer 0.8428355957776772
coast earth station 0.826395173453997
earth station antenna 0.62775263951735
earth station Hpa 0.555203619909502
earth station antennas 0.555203619909502
earth station receiver 0.518099547511312
earth station transmit 0.482503770739065
earth station equipment 0.471794871794872
earth station number 0.468174962292609
earth station identification 0.445399698340875
earth station HPàs 0.443891402714932
earth station complexes 0.442383107088989
earth station located 0.407541478129713
earth station owner 0.406033182503771
earth station receive 0.376470588235294
earth station transmitter 0.37420814479638

FIGURA 4. Llista de candidats a terme ordenats per valor de ponderació

5. CONCLUSIONS

La combinació de diverses estratègies estadístiques aplicada a l'extracció d'unitats pròpies d'un àmbit d'especialitat permet d'identificar amb més eficàcia aquest tipus d'unitats que no pas considerar els resultats obtinguts a partir d'una sola estratègia estadística. Els resultats que hem obtingut fins ara així ens ho confirmen; per aquest motiu, treballem per a poder identificar quina és la combinació de mesures estadístiques més adequada amb l'objectiu d'extreure un major nombre d'unitats terminològiques procedents de diferents corpus d'especialitat. I ho fem contrastant els resultats que ens ofereix cada mesura estadística amb els valors de freqüència, rang i *tf-idf*, tal com acabem de descriure.

En definitiva, el fet de poder identificar unitats terminològiques a partir d'un procés automatitzat facilita enormement l'elaboració de llistes de paraules clau i la construcció de taxonomies i futures ontologies en l'àmbit pròpiament de la documentació, i constitueix el material de partida per a poder plantejar un treball terminològic en el qual s'hagi de processar un gran volum de corpus en una llengua o en més d'una llengua.

6. REFERÈNCIES BIBLIOGRÀFIQUES

ARDANUY, J (2003). «Els models matemàtics de recuperació de la informació i la seva implementació en motors de cerca de propòsit general» [en línia]. A: *E-prints in Library and Information Science*. <http://eprints.rclis.org/archive/00007953/01/motors.pdf> [Consulta: 29 maig 2009].

BAEZA-YATES, R.; RIBEIRO-NETO, B (1999). *Modern information retrieval*. ACM Press.

BANERJEE, S.; PEDERSEN, T. (2003). «The Design, Implementation and Use of the Ngram Statistics Package» [en línia]. A: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mèxic, p. 370-381. <http://www.d.umn.edu/~tpederse/Pubs/cicling2003-2.pdf> [Consulta: 29 maig 2009].

CHURCH, K. W.; HANKS, P (1990). «Word association norms, mutual information and lexicography» [en línia]. *Computational Linguistics*, núm. 16, p. 22-29. <http://acl.ldc.upenn.edu/J/J90/J90-1003.pdf> [Consulta: 29 maig 2009].

CODINA, L.; ROVIRA, C (2002). «Information Retrieval Techniques» [en línia]. A: *Organización y recuperación de la información*. Universitat Oberta de Catalunya. (Documents de Lectura) <http://cv.uoc.es/cdocent/BOIQM7V2N6_PVI7JZGVG.pdf> [Consulta: 29 maig 2009].

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 85-92 DOI: 10.2436/15.2503.02.4

La documentació aplicada a la traducció jurídica

Eivor Jordà
Centre Universitari ESTEMA
València

Resum

Els terminòlegs i documentalistes s'ocupen de recopilar, descriure i catalogar informació, i els traductors especialitzats ens servim de moltes de les seves produccions: glossaris, corpus, bibliografies, catàlegs, etc. No obstant això, el traductor es veu constantment obligat a exercir de documentalista, ja que les seves necessitats informatives són molt variables. Així, cal que professionals adients transmetin als traductors no tant productes acabats sinó més aviat l'habilitat per a recuperar dades, és a dir, l'*alfabetització informacional*, tot i tenir en compte les particularitats de les diferents especilitats de la traducció, com ara la traducció jurídica.

Paraules clau: alfabetització informacional, documentació, informació, llenguatges d'especialitat, traducció especialitzada

Abstract: *Documentation applied to legal translation*

Terminologists and documentalists must compile, describe and catalogue information, and specialised translators make use of many of theirs products: glossaries, corpora, bibliographies, catalogues, etc. Nevertheless, the translator usually acts as a documentalist since his information needs are very changeable. Thus, suitable professionals should transfer to translators not so much finished products but the ability to data recovery, the so called *information literacy*, taking to account the special features of the different specialities in translation, as legal translation.

Key words: information literacy, documentation, information, specialised language, specialised translation.

1. La documentació aplicada a la traducció

Des de la meva perspectiva de traductora i docent especialitzada en la branca de la traducció jurídica, la documentació adquireix un interès fonamentalment pràctic com a mitjà per a solucionar problemes traductològics. Així doncs, la meva contribució a aquesta Jornada de «Terminologia i documentació» espero que sigui la de transmetre a terminòlegs i documentalistes les necessitats dels traductors jurídics en matèria de documentació des del punt de vista de l'usuari. Els terminòlegs i documentalistes s'ocupen de recopilar, descriure i catalogar informació amb la finalitat que aquesta informació resulti fàcilment recuperable per a qui la pugui necessitar. En aquest sentit, els traductors ens servim de moltes de les seves produccions: glossaris, corpus, bibliografies, catàlegs, etcètera.

No obstant això, el traductor es veu constantment obligat a exercir de documentalista. Això es deu al fet que les necessitats informatives dels traductors són molt variables i, per tant, resulta impossible delimitar-les amb la finalitat de realitzar algun tipus de compilació o sistematització prèvia. En conseqüència, per al traductor, la documentació cobra sentit com a mètode per a obtenir amb caràcter immediat informació puntual. En definitiva, el concepte de documentació al que em refereixo és el de l'habilitat per a recuperar dades, ja que, tal com Dora Sales (2006, p. 62) afirma: «Como usuario de la documentación, el traductor es selectivo y especializado. Lo que le interesa es saber cómo identificar, evaluar, utilizar y rentabilizar las fuentes de información requeridas para cubrir sus necesidades en cada momento».

Si amb anterioritat a l'era d'Internet, aquest tipus de documentació resultava complicat (encara que per motius molt diferents), actualment ens trobem amb el problema afegit de l'esmunyedís món dels continguts digitals. Tot això ha contribuït al que Ernest Abadal (2005, p. 32) denomina *desbordament cognitiu*, és a dir, un excés d'informació al qual s'afegeix, a més a més, el problema de la desorganització. El traductor professional s'ha de saber moure amb agilitat per aquest univers, ja que «[…] de la pertinencia y la calidad de las fuentes consultadas así como del tiempo empleado en acceder a la información dependerá en gran medida la calidad y rentabilidad de la traducción» (Rocío Palomares, 2000, p. 16). Amb aquesta finalitat, el traductor ha de saber localitzar, validar i utilitzar correctament totes les fonts d'informació al seu abast.

2. L'especificitat del llenguatge jurídic

L'especificitat de la traducció jurídica resideix bàsicament en el llenguatge jurídic com a reflex del sistema conceptual propi del dret. Per això, per a dominar

aquesta branca de la traducció, no n'hi ha prou amb el coneixement d'una segona llengua i una terminologia concreta, sinó que resulta indispensable, d'una banda, saber situar-se dintre de cada ordenament jurídic i, d'altra banda, conèixer els gèneres textuals pertanyents al camp del dret de les cultures en joc (cf. Pilar Blanco, 2003, p. 172). Aquesta és precisament una de les diferències bàsiques del llenguatge jurídic respecte d'altres llenguatges d'especialitat; mentre que molts d'aquests llenguatges tracten sobre matèries universals (sobretot en l'àmbit cientificotècnic), el llenguatge jurídic és en gran mesura de caràcter cultural, per la qual cosa, en molts casos, no existeixen equivalències conceptuals.

El llenguatge jurídic és, a més a més, extremament conservador en contraposició al caràcter dinàmic d'altres tecnolectes. En aquest sentit, podem contraposar el llenguatge científic, «[…] íntimamente vinculado al proceso de la denominada *creación científica*» (José López, 2000, p. 47), al jurídic, que està «[…] anclado en fórmulas arcaizantes y expresiones que permanecen invariables desde hace siglos» (Anabel Borja, 2000, p. 12). Un altre element diferenciador entre ambdós tipus de llenguatges especialitzats és la precisió dels termes cientificotècnics, respecte del vocabulari jurídic, en el qual preval la sinonímia i la polisèmia. Així doncs, en els textos jurídics és molt usual l'aparició de cadenes de sinònims l'objectiu dels quals és la matisació conceptual davant la vaguetat semàntica de la qual pequen molts termes jurídics. De la mateixa manera, és freqüent que en el vocabulari jurídic es doni també el fenomen de la polisèmia, ja que un mateix terme pot referir-se a diferents conceptes segons la branca del dret a la qual es fa referència.

3. Necessitats documentals de la traducció jurídica

Tenint en compte les característiques diferencials del llenguatge jurídic, Esther Monzó (2005, p. 137-141) planteja com a necessitats documentals específiques per a la traducció jurídica: la definició comparada de termes, la ubicació d'un concepte en el sistema jurídic al qual pertany, la detecció dels contextos originals, l'estructura potencial de gèneres originals en la llengua d'arribada, l'estructura potencial dels transgèneres (documents traduïts) en la llengua d'arribada i altres qüestions de caràcter estilístic. Per la meva banda, considero que, per a aquesta branca concreta de la traducció, els dubtes que apareixen pròpiament en el procés traductor podrien reduir-se a tres tipus: conceptuals, terminològics i fraseològics. En la pràctica traductològica és molt freqüent que la resolució d'aquests dubtes es realitzi de manera improvisada o intuïtiva. Per aquest motiu, l'esquema següent podria servir com a guia per a canalitzar les recerques en funció del tipus de dubte, l'objectiu que es persegueix i les fonts que s'haurien de consultar en cada cas.

Dubte	*Objectiu*	*Fonts per a resoldre'l*
Conceptual (en llengua A[1])	Determinar el significat exacte d'un terme en la llengua A.	1. Fonts especialitzades (en llengua A): enciclopèdies, manuals de dret, monografies de dret, articles de revistes de dret, compendis legislatius i jurisprudencials, etc. 2. Fonts directes: experts en una matèria concreta. 3. Fonts terminològiques (en llengua A): diccionaris monolingües o bilingües generals i especialitzats, lèxics, glossaris, bancs de dades terminològiques, etc.
Terminològic (en llengua B)	Localitzar les possibles equivalències d'aquest terme en la llengua B.	1. Fonts especialitzades (en llengua B): enciclopèdies, manuals de dret, monografies de dret, articles de revistes de dret, compendis legislatius i jurisprudencials, etc. 2. Fonts directes: experts en una matèria concreta. 3. Fonts terminològiques (en llengua B): diccionaris monolingües o bilingües generals i especialitzats, lèxics, glossaris, bancs de dades terminològiques, etc.
Conceptual (en llengua B)	Determinar el significat exacte dels diferents termes equivalents en la llengua B.	1. Fonts especialitzades (en llengua B): enciclopèdies, manuals de dret, monografies de dret, articles de revistes de dret, compendis legislatius i jurisprudencials, etc. 2. Fonts directes: experts en una matèria concreta. 3. Fonts terminològiques (en llengua B): diccionaris monolingües o bilingües generals i especialitzats, lèxics, glossaris, bancs de dades terminològiques, etc.
Fraseològic (en llengua B)	Determinar si una expressió s'utilitza en la llengua B (en el mateix context i amb el mateix significat que per al text A).	1. Textos comparables: textos en llengua A i en llengua B (tant originals com traduccions) que arriben al grau màxim de similitud al text que hem de traduir. 2. Textos paral·lels: textos similars al text que s'ha de traduir dels quals disposem tant de l'original com de la traducció.

1. Denominem, en aquest text, *llengua A* la llengua de partida d'una traducció i *llengua B*, la llengua d'arribada.

El traductor jurídic s'enfronta amb freqüència de manera inevitable amb termes de la llengua de partida que ignora o no domina amb precisió, ja que «[…] tener el conocimiento requerido para la traducción jurídica en la gran variedad de temas que puede depararle su futuro profesional resulta imposible e igualmente ineficiente, como lo sería para un jurista especializarse y ejercer en todas las áreas del Derecho» (Esther Monzó, 2005, p. 124) (i més encara si tenim en compte que en traducció sempre hi ha dos ordenaments jurídics en joc). El que hauria de fer el traductor en aquests casos és seguir una estratègia de recerca documental lògica i organitzada. Primer, comprendre el significat del terme en la llengua de partida (dubte conceptual en la llengua A); segon, localitzar les possibles accepcions del terme en la llengua d'arribada (dubte terminològic en la llengua B); tercer, comprovar el significat exacte de cadascuna de les accepcions en la llengua de partida, generalment ofertes pel diccionari bilingüe (dubte conceptual en la llengua B), i quart i últim, verificar quina d'aquestes és l'accepció correcta en la llengua d'arribada en funció del gènere textual.

Com es pot veure en el quadre, les fonts consultades per als diferents tipus de dubtes poden coincidir parcialment. No obstant això, és important remarcar que la perspectiva des de la qual es realitzen aquestes consultes és diferent en cada cas. Així doncs, quan, per exemple, es consulta un diccionari bilingüe per resoldre un dubte conceptual, el que interessa és la definició que en molts casos s'inclou en algunes entrades d'aquests diccionaris; en canvi, quan es consulta aquest mateix diccionari per un dubte terminològic, el que es busca és establir els termes equivalents en la llengua d'arribada. A més a més, els dubtes terminològics també es poden solucionar acudint a fonts especialitzades quan, per exemple, tenim clar un concepte i acudim directament a monografies, compendis legislatius o altres amb la finalitat de localitzar el terme utilitzat en la llengua d'arribada per a referir-s'hi.

L'esquema presentat és, en part, aplicable a qualsevol dubte traductològic relatiu a referents culturals. No obstant això, les característiques pròpies del llenguatge jurídic justifiquen la necessitat d'una documentació aplicada a aquesta branca concreta de la traducció. D'una banda, perquè el dret constitueix una àrea de coneixement molt específica marcada per la idiosincràsia dels ordenaments jurídics. D'altra banda, perquè les fonts que ha de consultar el traductor jurídic posseeixen a més a més unes particularitats concretes. Pensem, per exemple, en la legislació i la jurisprudència com a fonts de documentació tant conceptual com terminològica. De la mateixa manera, la destacada estandardització tant de l'estructura com de les expressions dels textos jurídics obliguen al traductor a documentar-se mitjançant textos comparables i paral·lels. En aquest sentit, els dubtes fraseològics sí que constitueixen una característica pròpia de la documentació per a la traducció jurídica, ja que en molts casos van més enllà de l'àmbit terminològic del que s'ocupen normalment els diccionaris.

El principal problema amb el qual es troba el traductor jurídic a l'hora de documentar-se resideix en la dificultat per a localitzar les fonts. En principi, existeix tot un seguit de fonts que podríem denominar *ortodoxes*, com són els diccionaris, enciclopèdies, formularis, compendis, etc. A manera d'exemple, per al parell de llengües anglès-espanyol, tenim diccionaris impresos de prestigi reconegut, com és el cas del *Diccionario de términos jurídicos inglés-español, Spanish-English* de Enrique Alcaraz Varó i Brian Hughes. Així mateix, són uns quants els grups d'investigació que es dediquen des de fa un cert temps a l'elaboració de corpus de textos jurídics en diversos idiomes (com el GITRAD de la Universitat Jaume I de Castelló). Malgrat això, el traductor jurídic es troba sovint amb la necessitat de recórrer a Internet per resoldre dubtes conceptuals, terminològics o fraseològics per als quals no troba resposta en les fonts que hem anomenat *ortodoxes*.

Per la seva banda, en el camp de la documentació, s'ha començat a parlar de l'*alfabetització informacional* o *alfabetització de la informació* que María Pinto (2005, p. 23) defineix com: «[…] el aprendizaje de habilidades, competencias, conocimientos y valores para el acceso, uso y comunicación de la información en cualquiera de sus formas, con el fin de generar profesionales y usuarios competentes, entrenados en el hábito de saber identificar y registrar las fuentes de la información, saber procesar y producir información propia, saber discriminar y valorar la información procesada y saber generar productos de comunicación de calidad». Al meu judici, aquesta és la clau de la documentació aplicada a la traducció. No obstant això, considero que s'hauria de donar un pas més enllà en el sentit de segmentar aquesta disciplina en funció de les particularitats de cada branca de traducció; ja que, com s'ha vist, els problemes de documentació són, en molts casos, diferents i les fonts, distintes.

4. Conclusió

Com a conclusió, voldria insistir en la necessitat que terminòlegs i documentalistes donin suport als traductors en la seva *alfabetització informacional*. En primer lloc, desenvolupant estratègies de resolució de dubtes (com la que he exposat) en les quals s'estableixi, d'una banda, la delimitació dels dubtes traductològics (quina informació necessito?) i, d'altra banda, el procediment a seguir per a resoldre'ls (com puc trobar aquesta informació?). En segon lloc, elaborant classificacions de fonts específiques per a cada branca de la traducció. En aquest cas, al traductor l'interessa, més que unes llistes de documents, que en poc temps queden obsoletes, una tipologia de fonts en funció de les especialitats de la traducció i dels tipus de dubtes. En tercer lloc, desenvolupant estratègies de recerca a Internet enfocades a la resolució de dubtes traductològics. Amb aquesta finalitat el documentalista ha de tenir present, d'una banda, la rapidesa amb la qual el traductor neces-

sita trobar respostes i, de l'altra, el grau de profunditat de la informació requerida. En aquest cas, la perspectiva del traductor i del documentalista o terminòleg és un poc distinta; el traductor no pot documentar-se de manera exhaustiva davant cadascun dels dubtes que se li plantegen, sinó que necessita trobar amb celeritat una solució a un problema de les diverses possibles.

No hem d'oblidar el que M. Teresa Cabré (2000, p. 35) afirma sobre la funció de la documentació: «El objetivo fundamental del trabajo documental es facilitar la recuperación de la información [...]». En aquest sentit, podríem dir que la documentació està al servei de les matèries a les quals s'aplica i, en conseqüència, l'èxit de la seva labor es podria mesurar pel grau d'eficiència en l'obtenció de la informació per part de qui la sol·licita. Així doncs, pel que fa a la documentació aplicada a la traducció jurídica, el que en aquest cas es demana és (utilitzant el símil de la solidaritat amb els països emergents) que els documentalistes no ens abasteixin d'aliments, sinó que ens ensenyin a cultivar la nostra pròpia terra. Es tracta d'una qüestió de cooperació necessària, en què el que els traductors necessiten són estratègies i eines de treball, i, per a dissenyar-les, els documentalistes han de conèixer l'ús que se'n farà. Com afirma Roberto Mayoral (1994, p. 118): «[...] el trabajo de traducción es en gran medida un problema de documentación».

5. Bibliografia

Abadal, Ernest (2005). «Contenidos digitales en Internet: algunos problemas». A: García del Toro, Cristina; García Izquierdo, Isabel (ed.). *Experiencias de traducción*: *reflexiones desde la práctica traductora*. Castelló: Universitat Jaume I, p. 31-42.

Alcaraz, Enrique; Hughes, Brian (2007). *Diccionario de términos jurídicos inglés-español, Spanish-English*. Barcelona: Ariel.

Blanco, Pilar (2003). «Problemas de la documentación jurídica». A: Vega Cernuda, Miguel Ángel (coord.). *Una mirada al taller de San Jerónimo*: *bibliografías, técnicas y reflexiones en torno a la traducción*. Madrid: Universidad Complutense de Madrid, p. 171-178.

Borja, Anabel (2000). *El texto jurídico inglés y su traducción al español*. Barcelona: Ariel.

Cabré, M. Teresa (2000). «Terminología y documentación». A: Gonzalo García, Consuelo; García Yebra, Valentín. *Documentación, terminología y traducción*. Madrid: Síntesis, p. 31-45.

López Yepes, José (2000). «Los investigadores como creadores de lenguaje científico: introducción al estudio terminológico de la *documentación* en España». A: Gonzalo García, Consuelo; García Yebra, Valentín. *Documentación, terminología y traducción*. Madrid: Síntesis, p. 45-60.

Mayoral Asensio, Roberto (1994). «La documentación en traducción». A: Jacoby, Lucien (ed.). *Traducción, interpretación, lenguaje*. Madrid: Actilibre, p. 107-118.

Monzó Nebot, Esther (2005). «Cómo traducir derecho sin ser jurista: nuevas fuentes y

métodos documentales para la traducción jurídica». A: SALES SALVADOR, Dora (ed.). *La biblioteca de Babel: documentarse para traducir*. Granada: Comares, p. 123-146.

PALOMARES PERRAUT, Rocío (2000). *Recursos documentales para el estudio de la traducción.* Màlaga: Universidad de Málaga.

PINTO MOLINA, María (2005). «Alfabetización en información para traductores: propuesta del modelo ALFINTRA». A: SALES SALVADOR, Dora (ed.). *La biblioteca de Babel: documentarse para traducir*. Granada: Comares, p. 19-32.

SALES SALVADOR, Dora (2006). *Documentación aplicada a la traducción: presente y futuro de una disciplina*. Gijón: Trea.

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 93-102 DOI: 10.2436/15.2503.02.5

El vocabulari de preservació i conservació del patrimoni documental

Maria Elvira
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

Resum

Aquesta comunicació presenta un vocabulari breu de termes de l'àmbit científic sobre preservació i conservació del patrimoni documental, recuperables a l'hora de produir o traduir un text d'aquest llenguatge d'especialitat. La llengua de referència és el català, però inclou també equivalències en anglès i en espanyol.

Paraules clau: documentació, patrimoni documental, traducció especialitzada, vocabulari.

Abstract: *The vocabulary on the preservation and conservation of documentary heritage*

This communication presents a brief vocabulary of scientific terms on the preservation and conservation of documentary heritage which may be retrieved when writing or translating texts on this specialist language. The reference language is Catalan, but it also includes equivalents in English and Spanish.

Key words: documentation, documentary heritage, specialised translation, vocabulary.

El vocabulari que ara es presenta, que és el producte d'un treball lent però llarg, es va iniciar en 2003 quan l'autora es va presentar a les oposicions per a la plaça de professor titular d'escola universitària i va elaborar com a part del material docent un glossari en espanyol de termes especialitzats. Posteriorment, ha recollit els termes nous que, en forma impresa o digital, ha trobat en la preparació de les classes de l'assignatura de preservació i conservació.

Es tracta d'un recull de gairebé dos-cents cinquanta termes catalans, de l'àmbit de la preservació i conservació, i dels seus equivalents en espanyol i en anglès. Té la voluntat de ser un vocabulari equivalent als d'arxivística i biblioteconomia impul-

sats per la Comissió de Dinamització Lingüística de la Facultat de Biblioteconomia i Documentació. Abasta la preservació tradicional i la preservació digital, totes dues de gran transcendència però de dificultat, cost i treball desiguals.

És un glossari adreçat principalment als professionals i als usuaris de les biblioteques i els arxius i també als estudiants de biblioteconomia i documentació. És un vocabulari professional que interessa també els no-professionals, cosa que no passa amb glossaris paral·lels, que només tenen utilitat per als professionals; i això perquè la conservació i, especialment, la preservació són un deure de totes les persones que tenen relació amb el patrimoni documental, perquè en són usuàries o responsables.

La preservació i la conservació no tenen encara una terminologia acceptada de manera genera, i per això aquesta comunicació pretén aportar una eina més amb la qual treballar en el procés necessari de normalització de la terminologia.

Glossaris utilitzats

Canals Arumí, M. Teresa; Gentile, Mónica E. *Glosario para restauradores de papel: español-catalán-inglés, català-anglès-espanyol, English-Catalan-Spanish*. Actas del V Congreso Nacional de Historia del Papel en España (Sarrià de Ter, 2-4 octubre 2003). Girona: CCG Ediciones: Ajuntament de Sarrià de Ter, 2003, p. 559-575.

National Library of Australia. *Library preservation glossary* [en línia]. <http://www.nla.gov.au/chg/gloss.html> [Consulta: 19 maig 2009].

Ruusalepp, Raivo. *AHDS Digital Preservation Glossary* [en línia]. Última versió, 2003. <http://ahds.ac.uk/exec/creating/glossary.htm> [Consulta: 19 maig 2009].

Vergara Peris, José. «Glosario». A: *Conservación y restauración de material cultural en archivos y bibliotecas*. 3a ed., renov. i ampl. València: Generalitat Valenciana. Conselleria d'Educació, Cultura i Esport, 2005, p. 222-238.

Vocabulari d'arxivística: català-castellà-anglès. Barcelona: Universitat de Barcelona. Serveis Lingüístics de la Universitat de Barcelona, 2005.

ANNEX

Vocabulari de preservació i conservació del patrimoni documental

Abreviatures

adj. adjectiu
f. nom femení singular
f. pl. nom femení plural
m. nom masculí singular
m. pl. nom masculí plural
n. nom singular
n. pl. nom plural
v. verb
v. tr. verb transitiu

Català	*Español*	*English*
abandó benigne *m.*	abandono benigno *m.*	benign neglect *n.*
abrasió *f.*	abrasión *f.*	abrasion *n.*
absorció *f.*	absorción *f.*	absorption *n.*
accessibilitat *f.*	accesibilidad *f.*	accessibility *n.*
acetat de cel·lulosa *m.*	acetato de celulosa *m.*	cellulose acetate *n.*
acetona *f.*	acetona *f.*	acetone *n.*
àcid *m.*	ácido *m.*	acid *n.*
àcid etílic *m.*	ácido etílico *m.*	ethylic acid *n.*
àcid fènic *m.*	ácido fénico *m.*	phenic acid *n.*
àcid oxàlic *m.*	ácido oxálico *m.*	oxalic acid *n.*
acidesa *f.*	acidez *f.*	acidity *n.*
acrílic *adj.*	acrílico *adj.*	acrylic *adj.*
adhesiu *m.*	adhesivo *m.*	adhesive *n.*
agent escumós *m.*	agente espumoso *m.*	soap agent *n.*
agent tensoactiu *m.*	agente tensoactivo *m.*	surfactant *n.*
aigua destil·lada *f.*	agua destilada *f.*	distilled water *n.*
aigua oxigenada *f.*	agua oxigenada *f.*	oxygenated water *n.*
aiguacuit *m.*	engrudo *m.*	paste *n.*
aire condicionat *m.*	aire acondicionado *m.*	air conditioning *n.*
alcalinitat *f.*	alcalinidad *f.*	alkalinity *n.*
alcohol etílic *m.*	alcohol etílico *m.*	etylic alcohol *n.*
aldehid fòrmic *m.*	aldehido fórmico *m.*	formic aldehyde *n.*
alfacel·lulosa *f.*	alfacelulosa *f.*	alpha cellulose *n.*
alum *m.*	alumbre *m.*	alum *n.*
aminobenzè *m.*	aminobenceno *m.*	aminobenzene *n.*
amoníac *m.*	amoniaco / amoníaco *m.*	ammonia *n.*
anilina *f.*	anilina *f.*	aniline *n.*

Català	*Español*	*English*
aplanar *v. tr.*	alisar *v. tr.*	flatten *v.*
argó *m.*	argón *m.*	argon *n.*
arqueologia digital *f.*	arqueología digital *f.*	digital archeology *n.*
arxiu *m.*	archivo *m.*	archive *n.*
atmosfera inerta *f.*	atmósfera inerte *f.*	inert atmosphere *n.*
autenticitat *f.*	autenticidad *f.*	authenticity *n.*
avaluació *f.*	evaluación *m.*	evaluation *n.*
bacteri *m.*	bacteria *f.*	bacterium *n.*
badana *f.*	badana *f.*	sheepskin *n.*
benzè *m.*	benceno *m.*	benzene *n.*
blanqueig *m.*	blanqueamiento *m.*	bleaching *n.*
blanquejant òptic *m.*	blanqueador óptico *m.*	optical bleach *n.*
bressol de llibre *m.*	cuna de libro *f.*	book cradle *n.*
calendari de conservació *m.*	calendario de conservación *m.*	conservation calendar *n.*
canvi de format *m.*	cambio de formato *m.*	format change *n.*
carbó actiu *m.*	carbón activo *m.*	activated carbon *n.*
carbonat càlcic *m.*	carbonato cálcico *m.*	calcium carbonate *n.*
càrrega *f.*	carga *f.*	load *n.*
cartró *m.*	cartón *m.*	cardboard *n.*
cera microcristal·lina *f.*	cera microcristalina *f.*	microcrystalline wax *n.*
cera natural *f.*	cera natural *f.*	natural wax *n.*
cinta adhesiva *f.*	cinta adhesiva *f.*	adhesive tape *n.*
cinta magnètica *f.*	cinta magnética *f.*	magnetic tape *n.*
clapat *adj.*	moteado *adj.*	foxing *adj.*
climatització *f.*	climatización *f.*	heating, ventilating and air conditioning system (HVAC) *n.*
cloramina T *f.*	cloramina T *f.*	chloramine-T *n.*
cloroform *m.*	cloroformo *m.*	chloroform *n.*
clorur de calç *m.*	cloruro de cal *m.*	bleaching powder *n.*
clorur de polivinil *m.*	cloruro de polivinilo *m.*	polyvinyl chloride *n.*
cola *f.*	cola *f.*	glue *n.*
cola animal *f.*	cola animal *f.*	animal glue *n.*
cola d'arròs *f.*	cola de arroz *f.*	rice glue *n.*
cola de peix *f.*	cola de pescado *f.*	fish glue *n.*
colofònia *f.*	colofonia *f.*	colophony *n.*
colorant àcid *m.*	colorante ácido *m.*	acid dye *n.*
component orgànic volàtil *m.*	componente orgánico volátil *m.*	volatile organic compound *n.*
comunitat designada d'usuaris *f.*	comunidad designada de usuarios *f.*	designated user community *n.*

Català	*Español*	*English*
confiança *f.*	confianza *f.*	trust *n.*
conservació *f.*	conservación *f.*	conservation *n.*
conservació en fred *f.*	conservación en frío *f.*	cold storage *n.*
conservació preventiva *f.*	conservación preventiva *f.*	preventive conservation *n.*
conservador *m.*	conservador *m.*	curator *n.*
consolidació *f.*	consolidación *f.*	consolidation *n.*
contaminació ambiental *f.*	contaminación ambiental *f.*	environmental contamination *n.*
contaminació atmosfèrica *f.*	contaminación atmosférica *f.*	air pollution *n.*
contaminant *m.*	contaminante *m.*	pollutant *n.*
contracció *f.*	contracción *f.*	shrinkage *n.*
còpia de preservació *f.*	copia de preservación *f.*	preservation copy *n.*
còpia de seguretat *f.*	copia de seguridad *f.*	back-up copy *n.*
corbament *m.*	alabeo *m.*	warping *n.*
dades digitals *f. pl.*	datos digitales *m. pl.*	digital data *n. pl.*
enregistrador de dades *m.*	registrador de datos *m.*	data logger *n.*
decoloració *f.*	decoloración *f.*	discoloration *n.*
degradació enzimàtica *f.*	degradación enzimática *f.*	enzymatic degradation *n.*
degradació fotoquímica *f.*	degradación fotoquímica *f.*	photochemical degradation *n.*
dipòsit institucional *m.*	depósito institucional *m.*	institutional repository *n.*
desacidificació *f.*	desacidificación *f.*	deacidification *n.*
deselecció *f.*	deselección *f.*	deselection *n.*
deshumitejar *v. tr.*	deshumidificar *v. tr.*	dehumidify *v.*
desinfecció *f.*	desinfección *f.*	disinfection *n.*
destrucció *f.*	destrucción *f.*	destruction *n.*
digitalització *f.*	digitalización *f.*	digitization *n.*
diòxid de nitrogen *m.*	dióxido de nitrógeno *m.*	nitrogen dioxide *n.*
diòxid de sofre *m.*	dióxido de azufre *m.*	sulfur dioxide *n.*
dipòsit *m.*	depósito *m.*	deposit *n.*
dipòsit legal *m.*	depósito legal *m.*	legal deposit *n.*
dipòsit voluntari *m.*	depósito voluntario *m.*	voluntary deposit *n.*
disc òptic *m.*	disco óptico *m.*	optical disc *n.*
document efímer *m.*	documento efímero *m.*	ephemeral document *n.*
drap *m.*	trapo *m.*	rag *n.*
drets d'autor *m. pl.*	derechos de autor *m. pl.*	copyright *n.*
durabilitat *f.*	durabilidad *f.*	durability *n.*
efecte d'hivernacle *m.*	efecto invernadero *m.*	greenhouse effect *n.*
elements essencials *m. pl.*	elementos esenciales *m. pl.*	essential elements *n. pl.*
eliminació *f.*	eliminación *f.*	elimination *n.*
emmagatzematge *m.*	almacenamiento *m.*	storage *n.*
emulació *f.*	emulación *f.*	emulation *n.*
emulsió *f.*	emulsión *f.*	emulsion *n.*

Català	*Español*	*English*
encapsulació *f.*	encapsulación *f.*	encapsulation *n.*
encolatge *m.*	encolado *m.*	sizing *n.*
encriptació *f.*	encriptación *f.*	encryption *n.*
engrut *m.*	engrudo *m.*	paste *n.*
enllaç d'hidrogen *m.*	enlace de hidrógeno *m.*	hydrogen bond *n.*
enquadernació *f.*	encuadernación *f.*	binding *n.*
envelliment accelerat *m.*	envejecimiento acelerado *m.*	accelerated ageing *n.*
esborrany *m.*	borrador *m.*	draft *n.*
escarabat *m.*	escarabajo *m.*	beetle *n.*
esgrogueïment *m.*	amarilleamiento *m.*	yellowing *n.*
esquinç *m.*	desgarro *m.*	tear *n.*
estabilitat química *f.*	estabilidad química *f.*	chemical stability *n.*
esterilització *f.*	esterilización *f.*	sterilization *n.*
estratègia de preservació *f.*	estrategia de preservación *f.*	preservation strategy *n.*
etanol *m.*	etanol *m.*	ethanol *n.*
èter *m.*	éter *m.*	ether *n.*
exempt d'àcid *adj.*	libre de ácido *adj.*	acid free *adj.*
externalització *f.*	externalización *f.*	outsourcing *n.*
fenilamina *f.*	fenilamina *f.*	phenylamine *n.*
fenol *m.*	fenol *m.*	phenol *n.*
feromona *f.*	feromona *f.*	pheromone *n.*
fiabilitat *f.*	fiabilidad *f.*	reliability *n.*
filigrana *f.*	filigrana *f.*	watermark *n.*
filtre *m.*	filtro *m.*	filter *n.*
fixador *m.*	fijador *m.*	fixative *n.*
floridura *f.*	moho *m.*	mildew *n.*
foc lent *m.*	fuego lento *m.*	low heat *n.*
fong *m.*	hongo *m.*	fungus *n.*
formaldehid *m.*	formaldehido *m.*	formaldehyde *n.*
format *m.*	formato *m.*	format *n.*
format d'accés *m.*	formato de acceso *m.*	access format *n.*
format de difusió *m.*	formato de difusió *m.*	diffusion format *n.*
formol *m.*	formol *m.*	formol *n.*
fotodegradació *f.*	fotodegradación *f.*	photodegradation *n.*
fotòmetre *m.*	fotómetro *m.*	photometer *n.*
fotooxidació *f.*	fotooxidació *f.*	photoxidation *n.*
friabilitat *f.*	friabilidad *f.*	brittleness *n.*
fumigació *f.*	fumigación *f.*	fumigation *n.*
gel de sílice *m.*	gel de sílice *m.*	silica gel *n.*
gofrat[1] *m.* (sobre cuir)	gofrado[1] *m.* (sobre piel)	embossing *n.*
gofrat[2] *m.* (sobre paper)	gofrado[2] *m.* (sobre papel)	corrugating *n.*

Català	*Español*	*English*
goma d'esborrar *f.*	goma de borrar *f.*	eraser *n.*
gramatge *m.*	gramaje *m.*	weight *n.*
guarda *f.*	guarda *f.*	endpaper *n.*
hidròlisi *f.*	hidrólisis *f.*	hydrolysis *n.*
higròmetre *m.*	higrómetro *m.*	hygrometer *n.*
higroscòpia *f.*	higroscopia *f.*	hygroscopy *n.*
higrotermògraf *m.*	higrotermógrafo *m.*	hygrothermograph *n.*
hipoclorit càlcic *m.*	hipoclorito cálcico *m.*	calcium hypochlorite *n.*
hipoclorit de sodi *m.*	hioclorito de sodio *m.*	sodium hypochlorite *n.*
humitat absoluta *f.*	humedad absoluta *f.*	absolute humidity *n.*
humitat relativa *f.*	humedad relativa *f.*	relative humidity *n.*
ignifugar *v. tr.*	ignifugar *v. tr.*	fireproof *v.*
incinerador *m.*	incinerador *m.*	incinerator *n.*
inflament *m.*	hinchamiento *m.*	swelling *n.*
inhibidor fungicida *m.*	inhibidor fungicida *m.*	fungicidal buffer *n.*
insecte bibliòfag *m.*	insecto bibliófago *m.*	bookworm *n.*
integritat *f.*	integridad *f.*	integrity *n.*
laminació *f.*	laminación *f.*	lamination *n.*
laminadora *f.*	laminadora *f.*	laminating machine *n.*
lignina *f.*	lignina *f.*	lignin *n.*
llapis de pH *m.*	lápiz de pH *m.*	archivist's pen *n.*
llibres friables *m. pl.*	libros friables *m. pl.*	brittle books *n. pl.*
lligall *m.*	legajo *m.*	file *n.*
llom *m.*	lomo *m.*	spine *n.*
llum *f.*	luz *f.*	light *n.*
longevitat digital *f.*	longevidad digital *m.*	digital longevity *n.*
lumen *m.*	lumen *m.*	lumen *n.*
lux *m.*	lux *m.*	lux *n.*
luxímetre *m.*	luxómetro *m.*	lux meter *n.*
marca d'aigua digital *f.*	marca de agua digital *f.*	digital watermark *n.*
material nascut digital *m.*	material nacido digital *m.*	born digital material *n.*
metadades *f. pl.*	metadatos *m. pl.*	metadata *n. pl.*
metanol *m.*	metanol *m.*	methanol *n.*
microfilm *m.*	microfilm *m.*	microfilm *n.*
microfilm de preservació *m.*	microfilm de preservación *m.*	preservation microfilm *n.*
microfilm de seguretat *m.*	microfilm de seguridad *m.*	safety microfilm *n.*
midó *m.*	almidón *m.*	starch *n.*
migració *f.*	migración *f.*	migration *n.*
model referencial *m.*	modelo referencial *m.*	reference model *n.*
mostreig *m.*	muestreo *m.*	sampling *n.*
museu informàtic *n.*	museo informático *m.*	computer museum *n.*

Català	*Español*	*English*
naftalina *f.*	naftalina *f.*	naphthalene *n.*
nervi *m.*	nervio *m.*	nerve *n.*
neutralització *f.*	neutralización *f.*	neutralization *n.*
nitrat de cel·lulosa *m.*	nitrato de celulosa *m.*	cellulose nitrate *n.*
obsolescència tecnològica *f.*	obsolescencia tecnológica *f.*	technological obsolescence *n.*
ondulació *f.*	ondulación *f.*	curling *n.*
oxidació *f.*	oxidación *f.*	oxidation *n.*
ozó *m.*	ozono *m.*	ozone *n.*
panerola *f.*	cucaracha *f.*	cockroach *n.*
paper *m.*	papel *m.*	paper *n.*
paper alcalí *m.*	papel alcalino *m.*	alkaline paper *n.*
paper barrera *m.*	papel barrera *m.*	barrier paper *n.*
paper carbó *m.*	papel carbón *m.*	carbon paper *n.*
paper estucat *m.*	papel estucado *m.*	coated paper *n.*
paper japonès *m.*	papel japonés *m.*	Japanese paper *n.*
paper jaspiat *m.*	papel jaspeado *m.*	mottled paper *n.*
paper neutre *m.*	papel neutro *m.*	neutral paper *n.*
paper permanent *m.*	papel permanente *m.*	permanent paper *n.*
paper trencadís *m.*	papel quebradizo *m.*	brittle paper *n.*
paràsits *m. pl.*	parásitos *m. pl.*	parasites *n. pl.*
parimoni digital *m.*	patrimonio digital *m.*	digital heritage *n.*
pasta *f.*	pasta *f.*	pulp *n.*
pasta mecànica *f.*	pasta mecánica *f.*	mechanical pulp *n.*
pasta química *f.*	pasta química *f.*	chemical pulp *n.*
peix de plata *m.*	pececillo de plata *m.*	silverfish *n.*
potencial d'hidrogen (pH) *m.*	potencial de hidrógeno (pH) *m.*	potential of hidrogene (pH) *n.*
pla d'emergència *m.*	plan de emergencia *m.*	emergency plan *n.*
pla de desastre *m.*	plan de desastre *m.*	disaster plan *n.*
polièster *m.*	poliéster *m.*	polyester *n.*
polietilè *m.*	polietileno *m.*	polyethylene *n.*
polímer *m.*	polímero *m.*	polymer *n.*
polimerització *f.*	polimerización *f.*	polymerization *n.*
polipropilè *m.*	polipropileno *m.*	polypropylene *n.*
poll dels llibres *m.*	piojo de los libros *m.*	booklouse *n.*
preservació *f.*	preservacion *f.*	preservation *n.*
preservació de recursos digitals *f.*	preservación de recursos digitales *f.*	preservation of digital resources *n.*
preservació digital *f.*	preservación digital *f.*	digital preservation *n.*
programari *m.*	software *m.*	software *n.*
propietat significativa *f.*	propiedad significativa *f.*	significant property *n.*

Català	*Español*	*English*
protocol *m.*	protocolo *m.*	protocol *n.*
qualitat arxivística *f.*	calidad archivística *f.*	archival quality *n.*
qualitat d'arxiu *f.*	calidad archivo *f.*	archival quality *n.*
radiació infraroja *f.*	radiación infrarroja *f.*	infrared radiation *n.*
radiació ultraviolada *f.*	radiación ultravioleta *f.*	ultraviolet radiation *n.*
recurs digital *m.*	recurso digital *m.*	digital resource *n.*
reformatació *f.*	reformateado *m.*	reformatting *n.*
refrescament *m.*	refresco / refrescamiento *m.*	refreshing *n.*
registre *m.*	registro *m.*	record *n.*
registre dels formats digitals *m.*	registro de formatos digitales *m.*	digital format registry *n.*
registre digital *m.*	registro digital *m.*	digital record *n.*
reintegració *f.*	reintegración *f.*	reintegration *n.*
repositori *m.*	repositorio *m.*	repository *n.*
reserva alcalina *f.*	reserva alcalina *f.*	alkaline reserve *n.*
restauració *f.*	restauración *f.*	restoration *n.*
retenció *f.*	retención *f.*	retention *n.*
revisió *f.*	revisión *f.*	revision *n.*
roba *f.*	tela *f.*	cloth *n.*
rosegador *m.*	roedor *m.*	rodent *n.*
sabata de llibre *f.*	zapato para libro *m.*	book shoe *n.*
segellament digital de temps *m.*	sellado digital de tiempo *m.*	digital time stamp *n.*
selecció *f.*	selección *m.*	selection *n.*
signatura digital *f.*	firma digital *f.*	digital signature *n.*
sistema digital *m.*	sistema digital *m.*	digital system *n.*
Sistema Obert d'Arxivament d'Informació (OAIS) *m.*	Sistema Abierto de Archivado de Información (OAIS) *m.*	Open Archival Information System (OAIS) *n.*
sostenibilitat econòmica *f.*	sostenibilidad económica *f.*	economic sustainability n
suport digital *m.*	soporte digital *m.*	digital media *n.*
taca *f.*	mancha *f.*	stain *n.*
temperatura *f.*	temperatura *f.*	temperature *n.*
termita *f.*	termita *f.*	termite *n.*
timol *m.*	timol *m.*	thymol *n.*
tint *m.*	tinte *m.*	dye *n.*
tinta *f.*	tinta *f.*	ink *n.*
tinta cal·ligràfica *f.*	tinta caligráfica *f.*	calligraphy ink *n.*
tinta d'impremta *f.*	tinta de imprenta *f.*	printing ink *n.*
ultraviolímetre *m.*	ultraviolímetro *m.*	ultraviolet meter *n.*
validació *f.*	validación *f.*	validation *n.*

Català	*Español*	*English*
valor intrínsec *m.*	valor intrínseco *m.*	intrinsic value *n.*
versió analògica *f.*	versión analógica *f.*	analogic version *n.*
versió de preservació *f.*	versión de preservación *f.*	preservation version *n.*
versió digital *f.*	versión digital *f.*	digital version *n.*
viabilitat institucional *f.*	viabilidad institucional *f.*	institutional viability *n.*
vidriol *m.*	vitriolo *m.*	vitriol *n.*
virus informàtic *m.*	virus informático *m.*	computer virus *n.*
vitel·la *f.*	vitela *f.*	vellum *n.*

Sessió II

Ponència

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 105-116 DOI: 10.2436/15.2503.02.6

El futur de la informació acadèmica: Web semàntic / Web social, o tots dos?

Lluís Codina
Universitat Pompeu Fabra
Barcelona

Resum

Alguns professionals, com els periodistes, els traductors, els terminòlegs, els acadèmics, etc., requereixen una informació intensiva. Però aquesta informació no sol ser fàcil d'obtenir, ni de processar ni de recuperar més endavant. Aquesta ponència detalla quina ajuda poden oferir als anomenats *professionals intensius en informació* les diferents onades d'innovació del Web: Web 2.0 (Web social), Web 3.0 i Web semàntic. D'entrada, es proporcionen criteris diferenciadors entre ells i, finalment, s'explica l'impacte que poden tenir en els sistemes d'informació acadèmics.

Paraules clau: informació acadèmica, professionals intensius en informació, Web 2.0, Web 3.0, Web semàntic, Web social.

Abstract: *The future of academic information: Semantic Web / Social Web, or both?*

Some professionals, such as journalists, translators, terminologists, academics, etc., require intensive information. However, this information is not usually easy to find, process or subsequently retrieve. This paper details what help the different waves of innovation from the Web can offer the so-called information-intensive professionals: Web 2.0 (social Web), Web 3.0 and semantic Web. Initially, criteria to distinguish between them are proposed, and finally their possible impact on academic information systems is explained.

Key words: academic information, information-intensive professionals, Web 2.0, Web 3.0, semantic Web, social Web.

1. Ambients intensius en informació

Una característica d'alguns professionals és la necessitat constant de processar un cert tipus d'informació en el sentit més ampli de la paraula: cerca, descobri-

ment, anàlisi, emmagatzematge, recuperació, explotació, etc. Aleshores, diem que són professionals que desenvolupen la seva feina en ambients intensius en informació.

Economistes, juristes, periodistes, traductors, terminòlegs, enginyers i acadèmics en general (professors, estudiants, investigadors) són només alguns exemples d'aquesta classe de professions. Tota classe de professions gestionen o manipulen informació d'algun tipus. Per exemple, la persona que està al càrrec de la recepció d'un hotel, treballa bàsicament amb informació: dades de clients i de reserves, atendre consultes d'aquests sobre el mateix hotel, la ciutat, etc. En canvi, no totes les professions són intensives en informació.

La diferència és que en aquestes últimes la informació rellevant no sempre resulta fàcil ni d'obtenir ni de processar. De fet, la primera dificultat dels ambients intensius en informació consisteix en la necessitat de discriminar, entre grans volums d'informació, aquella petita fracció relativa que és realment útil a cada moment.

Sense voler treure importància a cap conjunt de professions, ja que sabem perfectament que totes són necessàries, el cert és que hi ha grans diferències en el sentit que ens interessa aquí. Si seguim amb l'exemple de la persona al càrrec d'una recepció, podem veure que, en general, no tindrà una gran dificultat per a accedir a la informació necessària per al seu treball: el client arribarà al taulell amb les dades de la seva reserva, que la persona de recepció contrastarà amb la informació de la bases de dades de l'hotel, etc. Treballa amb informació; però, per definició, la informació que necessita està molt ben delimitada i sempre o quasi sempre al seu abast amb un mínim esforç.

En canvi, imaginem un advocat que ha de cercar informació per a poder dur més bé la defensa del seu client, o un acadèmic que ha iniciat una nova línia de recerca, o un estudiant de doctorat que està fent la recerca per a la tesi doctoral.

En aquests últims contextos, que anomenem *intensius en informació*, el primer problema consisteix, sovint, que ni tan sols el primer element de la cadena, la necessitat d'informació, és fàcil de definir. Comparem aquestes dues necessitats d'informació: 1) «necessito saber quina és l'habitació de l'hotel que correspon al client amb la reserva X»; 2) «necessito saber quines son les polítiques més adequades per al desenvolupament econòmic sostenible».

Per a satisfer la necessitat d'informació núm. 1, només cal introduir el número de la reserva en la base de dades de l'hotel, i n'obtenim la resposta. Per a satisfer la necessitat d'informació núm. 2, el primer problema és identificar el sistema d'informació (si és que existeix) mitjançant el qual s'ha d'intentar aconseguir informació; el segon problema és com hem d'utilitzar el sistema d'informació mateix, amb quin llenguatge hem de formular la consulta, etc. A més, no hi ha mai un punt final: una informació obtinguda, en comptes de tancar el procés, pot obrir

nous interrogants i, per tant, la necessitat d'obrir una nova operació de cerca. Després, la informació obtinguda serà d'un tipus que pot distar molt de ser trivial o fàcil d'interpretar o assimilar. Típicament, consistirà en un conjunt d'informes més o menys complexos, tal vegada en una llengua estrangera, o en articles de revistes acadèmiques, potser amb punts de vista contraposats, etc. Finalment, el nostre professional necessitarà establir una manera mitjançant la qual en el futur pugui recuperar aquestes informacions per tal de reproduir dades i de poder-les citar.

Aportar solucions fiables als professionals dels ambients intensius en informació ha estat, alhora, l'objecte de diverses disciplines i de diversos sectors econòmics i empresarials. En el context de les universitats, les biblioteques han estat algunes de les eines utilitzades. En el món de l'empresa, els diversos sistemes d'informació corporatius n'han estat unes altres.

Tal com hem intentat argumentar, en aquests ambients no es útil qualsevol informació. Al contrari, necessitem:

— la millor informació;
— obtenir la informació en el moment oportú;
— amb costos assumibles.

Examinem breument aquests requeriments.

La millor informació: necessitem informació que sigui fiable, rellevant i que tingui l'orientació, el gènere, la morfologia i el format adequats.

En el moment oportú: descobrir un bona informació un cop finalitzada la tesi doctoral, un cop enviat l'article a la revista o un cop tancat el termini per a presentar la documentació d'un projecte, òbviament no serveix de res.

Amb costos assumibles: de temps, de diners i de processament. És evident que cap projecte no té ni recursos econòmics ni temps il·limitats. Per tant, la informació a obtenir s'ha d'ajustar a aquests paràmetres.

Des dels anys noranta tenim en el Web el sistema d'informació més formidable i més potent que mai no hauria pogut somiar la humanitat. El problema és que al Web hi ha tanta informació i amb una varietat tan gran de qualitat (des de la simple intoxicació a la millor tesi doctoral) que aquesta abundància d'informació és un problema en si mateixa. En els apartats següents intentarem presentar una panoràmica sobre què es preveu que serà el Web dels propers anys, pel que fa als sistemes que intenten proporcionar solucions als professionals intensius en informació.

2. Les tres onades del Web

L'expressió *Web 2.0* va tenir com a data oficial de naixement una conferència del mateix nom celebrada l'any 2004 als EUA, la qual cosa significa que, per aquelles dates, ja tenia un cert temps de vida. Únicament calia que algú identifiqués el canvi (o la tendència de canvi).

Sigui com sigui, el Web 2.0 ha resultat, *de facto*, una magnífica formula d'agitació cultural. Molt probablement, sense els canvis que ens ha aportat, el Web no tindria, ni de bon tros, l'abast gairebé universal que el caracteritza actualment.

A continuació presentarem una proposta d'identificació de components i de conceptualització del Web social (o Web 2.0), del Web 3.0 i del Web semàntic. És a dir, de les tres grans onades d'innovació del Web dels darrers lustres, que, malgrat que tendeixen a una més que lògica confluència no són exactament el mateix. La nostra proposta intentarà proporcionar criteris diferenciadors a partir d'identificar un petit, però probablement significatiu conjunt de característiques essencials de cadascuna de les tres onades. Posteriorment, presentarem de manera sintètica el possible impacte que tenen o que podrien tenir en les publicacions digitals i, sobretot, en els sistemes d'informació acadèmics.

2.1. *Web social i Web 2.0*

Atès que una de les característiques més importants del Web 2.0 és el seu fort component social (p. ex., continguts creats mitjançant «intel·ligència social»[1]), sovint el Web 2.0 és anomenat també *Web social*. El concepte de Web social és més ampli i alhora més concret[2] que el de Web 2.0, però per tal de simplificar la nostra exposició, a partir d'ara, si no indiquem el contrari, quan ens referirem al Web 2.0 ens referirem també al Web social.

En aquest sentit, i en relació amb l'expressió concreta de Web 2.0, una apreciació que ens sembla errònia sobre el Web actual consisteix a creure que ara «som» en el Web 2.0, tal com abans —en la dècada dels noranta per dir-ho així— se suposa que «érem» en el Web 1.0.

En realitat, totes les eres o versions del Web conviuen en el Web actual. El motiu és simple: una part molt important, possiblement la majoria, de les pàgines i documents del Web continuen «sent» Web 1.0, és a dir, pàgines i documents estàtics publicats pels gestors i responsables dels llocs web respectius. Al mateix temps, com sabem, cada cop més llocs estan incorporant elements del Web 2.0 i d'altres elements del que es considera el futur, a saber, elements del Web 3.0 o del Web semàntic.

En tot cas, entenem que hi ha un consens, més factual que no pas teoritzat de

1. Dos exemples notables d'aquesta anomenada «intel·ligència social» o «intel·ligència col·lectiva» serien la Viquipèdia i els sistemes basats en recomanacions com Digg o Technorati.

2. És més ampli perquè la característica social del Web 2.0 segurament transcendirà la mateixa Web 2.0 i formarà part dels futurs webs; i és més concret perquè, tot i la seva importància fonamental, el Web 2.0 té més components.

manera explícita o formal, en el sentit que el Web 2.0 estaria caracteritzat pels components principals següents:

1) Continguts creats pels usuaris. Els usuaris entesos com a «prossumidors» (productors i consumidors a la vegada).

2) Xarxes socials. El Web com a plataforma de relació social, personal i/o professional.

3) Aplicacions en línia. El Web com a plataforma per a executar aplicacions sense necessitat d'instal·lar programari addicional en l'ordinador.

4) Eines de col·laboració. Un cop més, el Web com a plataforma, en aquest cas per a donar suport i proporcionar eines a grups de treball que cooperen en una mateixa tasca o objectiu.

La figura 1 pretén reflectir aquestes idees d'una manera gràfica, i hi afegeix una consideració addicional: els quatre components estan vinculats d'una manera indissoluble entre si. Sense les aplicacions en línia, difícilment tindríem fenòmens

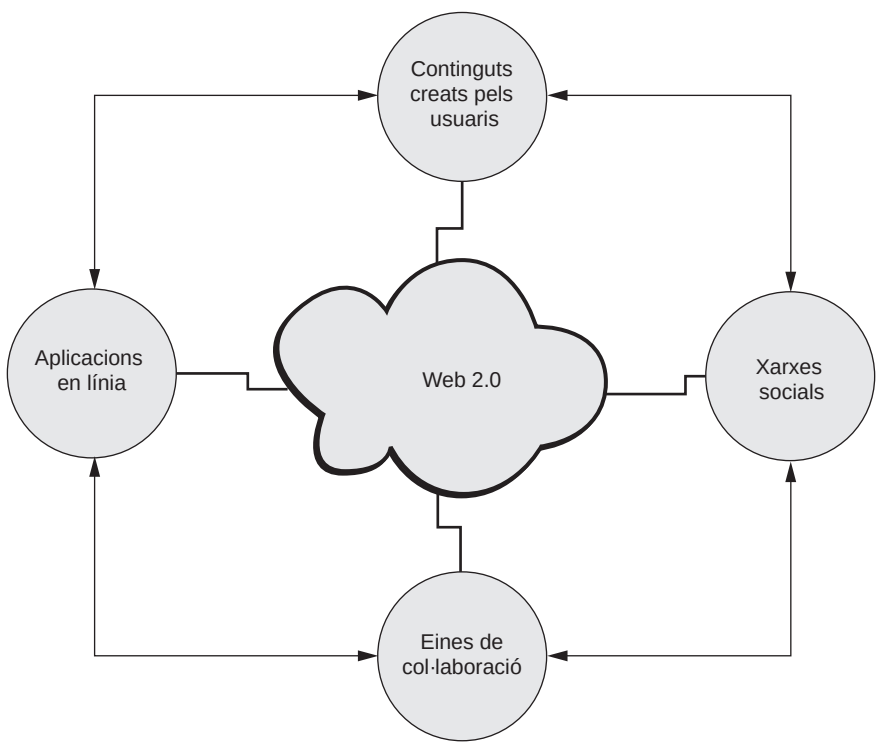


FIGURA 1. Components del Web 2.0

com la Viquipèdia (o la blogosfera en general); mentre que les eines de col·laboració no són més que un cas de computació en línia. Finalment, les xarxes socials faciliten i fomenten la distribució de continguts socials, i fan que tot es comporti com un sistema autosostingut.

2.2. *Web semàntic*

A final dels noranta es va iniciar un nou canvi al Web. Era un canvi, d'una banda, totalment independent del Web 2.0 i, alhora, més complex i molt més ambiciós. També —s'ha de dir— molt més «utòpic» o, si es vol, molt més vinculat amb una visió, i no el resultat d'una evolució natural. Es tracta del projecte del Web semàntic.

Aquí tenim una diferència important: el Web 2.0 és un resultat *de facto*. El Web 2.0 és com és i té les característiques que té sense que ningú hagi dissenyat aquests canvis de manera específica. En canvi, el Web semàntic sí que és el resultat d'un disseny. És un projecte conscient i dirigit, i no un simple (o complex) resultat de les coses, com en el cas anterior.

Concretament, el Web semàntic és el nom d'un projecte concebut, dissenyat, promogut i dirigit, almenys en els seus trets principals, pel Consorci World Wide Web (W3C). Com és sabut, aquest Consorci és el principal organisme de normalització i, a la vegada, un dels principals responsables de la dinamització del Web.

El director del Consorci, Sir Tim Berners-Lee, va ser el creador del Web i del llenguatge (X)HTML, que ha fet possible tant el Web d'«abans» com el d'«ara». La qüestió és que, a final dels noranta —tal com hem assenyalat—, Berners-Lee va considerar que el Web requeria canvis en profunditat i va llançar el projecte del Web semàntic. Actualment, uns deu anys després del seu llançament oficial, el projecte ha avançat molt poc; si més no, comparat amb les previsions inicials, que ara podem dir que van ser clarament visionàries. Aquestes previsions estaven vinculades amb unes perspectives més pròpies d'intel·ligència artificial que amb les possibilitats reals de les ciències de la computació. Fins i tot s'allunyen de la mateixa intel·ligència artificial actual, entesa com a disciplina científica i no pas com la barreja de ciència i pseudociència visionària que va ser entre la dècada dels seixanta i la dels vuitanta (quan dia sí dia també s'assegurava que l'any següent tindríem ordinadors intel·ligents).

Afortunadament, el projecte ha estat capaç de desenvolupar un conjunt de normes, llenguatges i tecnologies que estan tenint una influència positiva en el Web. Un altre efecte favorable del projecte del Web semàntic és que ha aconseguit una gran mobilització d'esforços, científics, empresarials i acadèmics, al voltant de l'objectiu d'un web molt més fàcil de utilitzar, i ha contribuït a atorgar una vida nova a algunes disciplines clàssiques que havien entrat en una mena d'estat en

suspensió amb el primer Web, com els llenguatges documentals o les ontologies. Entre els components conceptuals (no oblidem que és un projecte) més importants, podem assenyalar els següents:

— El Web entès com una gran base de dades. La idea és aconseguir que els documents publicats en el Web estiguin marcats de manera que siguin similars als registres d'una base de dades.

— Metadades. Els llocs web estarien caracteritzats per l'ús intensiu de sistemes de metadades com a part del seu marcatge.

— Ontologies i lògica formal. Es desenvoluparan ontologies per tal que els ordinadors interpretin la semàntica de les pàgines web, i sistemes de raonament automàtics basats en lògica formal que podran fer inferències.

— Agents d'usuari. Seran sistemes informàtics capaços de representar els interessos dels seus usuaris i d'interactuar amb altres sistemes sense intervenció dels mateixos usuaris.

Per tant, històricament, l'objectiu fundacional del Web semàntic va consistir a desenvolupar un complex de tecnologies que haurien de permetre als ordinadors, mitjançant l'ús d'agents d'usuari similars als navegadors actuals, no solament «entendre» el contingut de les pàgines, sinó també dur a terme raonaments sobre aquest contingut. La idea era aconseguir que l'enorme potencial real de coneixement registrat en documents es pogués interpretar com ho faria un ésser humà.

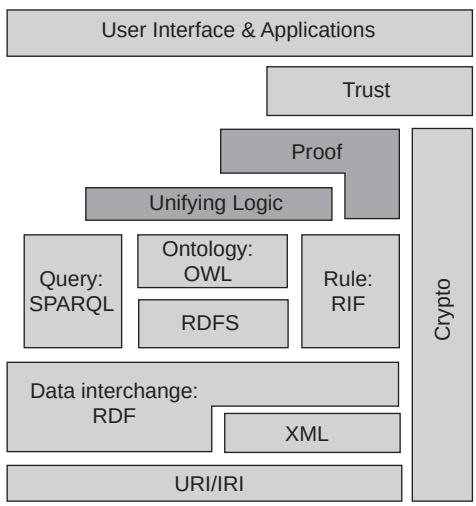


FIGURA 2. Diagrama del Web en forma de capes i mòduls (font: W3C, *http://www.w3.org*)

El diagrama anterior (figura 2) mostra, en forma de capes successives i de mòduls relativament autònoms, els components tecnològics i lògics principals del projecte. Es pot dir que s'han desenvolupat prou les tres capes inferiors: URI/IRI, XML i RDF. A grans trets, aquestes capes formen la infraestructura del Web actual, a banda de l'RDF, que és un sistema molt sofisticat de codificació de metadades encara molt poc implantat, sens dubte per la complexitat que té, però també pels escassos al·licients que n'aporta actualment l'ús.

El sistema URI/IRI és a la base del sistema d'adreces que identifica de manera única cada recurs del Web, bé sigui una pàgina web o qualsevol classe de document o d'objecte multimèdia. Per la seva banda, és difícil exagerar a hores d'ara la importància del llenguatge XML, que s'ha introduït amb una força enorme, no solament en el món del Web, sinó també en el món de l'ofimàtica i dels sistemes d'informació en general.

Amb el temps, el projecte del Web semàntic ha anant modificant els seus objectius, i en els últims anys s'ha centrat en aspectes molt més pragmàtics i realistes, tal com els que ha aconseguit, sense necessitat d'una direcció central, el Web 2.0 i tal com sembla que pot aconseguir la futura i hipotètica Web 3.0, que examinem a continuació.

2.3. *Web 3.0*

Fer servir números per a identificar generacions (o «onades», com diem aquí) del Web sembla una bona idea si l'hem de jutjar per l'èxit que va tenir la denominació *2.0*. No obstant això, mantenir aquest mètode sembla que està duent a una certa confusió. El fet és que encara no sabem què és o què podria ser el Web 3.0 i ja sovintegen les especulacions al voltant d'un suposat Web 4.0. No sembla gaire racional que cada analista que creu detectar algun canvi es llenci a posar un número més al Web, especialment perquè aquest mètode no requereix justificació (justificar una denominació, per alguna raó és molt més difícil) i, per tant, tot pot ser una mica (o molt) arbitrari.

La qüestió és que, malgrat tot, sembla que hi ha bases de canvi suficients per a pensar que som a l'inici d'una nova generació del Web, que podem anomenar com vulguem, és clar, però tot apunta que la denominació *3.0* ja és inevitable. Pel que fa a l'origen, sembla que la primera menció a un suposat Web 3.0 correspon a un article publicat en la influent publicació digital *ZDNet* del novembre del 2005 per Phil Wainewright.

Quins serien els trets d'aquest futur web? Aquí entrem en un terreny molt més difícil que en el cas dels dos webs anteriors, atès que no és ben bé ni un projecte dirigit (com el Web semàntic) ni un fet consumat (com el Web 2.0), sinó únicament una especulació més o menys solvent i més o menys basada en alguns ca-

sos aïllats (per més que siguin notables). Alguns analistes solucionen el (pseudo)problema identificant, sense més ni més, *Web semàntic* amb *Web 3.0*. D'aquesta manera, una manera de tancar —en fals— la discussió consistiria a prendre seriosament aquesta identificació.

No obstant això, alguns trets del Web dels últims anys són genuïns: no són ni del Web 2.0 ni estaven previstos ni deriven del Web semàntic. En relació amb aquest últim, el Web 3.0 comparteix en part amb el 2.0 un cert caràcter «espontani» o *de facto*.

Quines serien, malgrat tot, les característiques d'aquest nou Web? Nosaltres proposem les següents:

a) Computació en línia (*cloud computing*) i vinculació de dades i d'aplicacions. La computació en línia seria el següent pas lògic a les aplicacions en línia; p. ex., ara podem pensar en sistemes operatius en línia i en una computació íntegrament basada en el Web com a plataforma.

b) Agents d'usuari. Aquesta seria una de les característiques que vincula més el Web 3.0 amb el Web semàntic, ja que la idea és exactament la mateixa, tot i que molt més pragmàtica.

c) Amplada de banda. Comparats amb els trets anteriors, aquest sembla molt prosaic, però el cert és que l'augment constant de l'amplada de banda en el Web ens està conduint a un web que abans no hauria estat possible, com la comunicació audiovisual en directe o la computació en línia del primer punt.

d) Ubiqüitat del Web. Cada vegada més, el Web és omnipresent, i aquesta és una tendència creixent que ha donat pas a tot un web nou: el *Web mòbil*.

Pel que fa als punts anteriors, la vinculació de dades significa que cada vegada hi haurà més serveis d'informació que seran capaços d'agregar dades procedents de desenes o de centenars de fonts diferents i de mostrar-les als usuaris d'una manera tan unificada com si sempre haguessin estat perfectament unides («sense costures», com diuen els anglosaxons). Un exemple d'això podria ser el cercador Kosmix, la versió Glue del cercador Yahoo o les darreres versions del servei Google News.

La vinculació d'aplicacions seria també un altre pas en la línia de combinar les prestacions o les funcions de diversos programes per aconseguir resultats nous. Un exemple seria tant el mateix Google Maps, com l'ús que se'n fa dins d'altres aplicacions que, a la vegada, formen part de diversos serveis d'informació dins de pàgines web. En la mateixa línia, Google Earth seria un altre exemple de vinculació de diferents aplicacions i dades dins d'un sistema d'informació aparentment homogeni, capaç d'oferir una informació integrada sobre la totalitat del nostre planeta d'una manera que mai no hauríem somiat.

Per la seva banda, la idea dels agents d'usuari és la més especulativa i és alhora la que presenta una vinculació més forta amb el Web semàntic. Es tractaria, hipotèticament, d'una nova generació de navegadors o d'una nova generació de

plug-ins que podrien, d'alguna manera, explotar el contingut semàntic de les pàgines web amb capacitat de respondre d'una manera similar a com ho faria un ésser intel·ligent, a les preguntes o a les necessitats d'informació dels usuaris. En el cas extrem, aquests agents d'usuari desplegarien, fins i tot, capacitats de gestió en favor dels seus usuaris; per exemple, des de reservar uns seients en una funció de teatre fins a planificar una ruta, adquirir bitllets d'avió i contractar les reserves d'hotel d'un viatge a través de diversos països, etcètera.

3. Conclusions

El Web 2.0 ha tingut un impacte considerable en el que podríem anomenar «cibermitjans» (*social media*), és a dir, el complex format pels mitjans de comunicació en línia i el conjunt de nous mitjans socials, tals com YouTube, Flickr o la blogosfera en general. Ara bé, ha tingut un impacte menor en els sistemes de cerca com Google o Yahoo; ja que, de moment, les versions tipus Web 2.0 d'aquests cercadors (Google i Yahoo Glue) sembla que no estan adquirint una popularitat comparable a les versions estàndard.

Pel que fa al Web 3.0, és gairebé segur que l'impacte serà molt alt en el Web en general i en generarà un de nou, molt diferent de l'anterior. No obstant això, com sol succeir amb les tecnologies que triomfen de debò, ho farà d'una manera gairebé invisible o transparent. La qüestió és que, en el futur, serà rutinari fer servir serveis d'informació que presentaran respostes a les nostres preguntes combinant aplicacions i fonts d'informació molt diverses; però no en forma d'un llistat amb documents procedents de fonts heterogènies, sinó en forma de pàgines de resultats que semblaran documents unitaris amb la resposta (possibles respostes) presentada de manera directa.

És molt més dubtós l'impacte real del Web semàntic (més enllà de la indubtable influència acadèmica), en gran part perquè el programa màxim del Web semàntic està massa vinculat a la intel·ligència artificial. Tot i això, el Web semàntic pot tenir un bon paper com a proveïdor de llenguatges i estàndards per a facilitar la vinculació de dades i d'aplicacions del Web 3.0. D'aquesta manera, en la mesura que s'acabi fent realitat la fusió/identificació Web semàntic = Web 3.0, pot passar que el Web semàntic tingui èxit per una via mai no imaginada pels impulsors originaris.

En canvi, tant el Web 2.0 com el Web 3.0 poden tenir un gran impacte en els sistemes de gestió de la informació personal o PIM (*personal information managers*). La realitat és que l'abundància actual de fonts i de sistemes d'informació que se superposen parcialment, d'alternatives diferents, etc., fa que, d'una banda, sigui més fàcil que mai trobar informació, però, de l'altra, més complicat que mai organitzar-la d'una manera eficaç.

Dit d'una altra manera, actualment no és un problema trobar informació, sinó organitzar-la per tal que després puguem explotar-la i reutilitzar-la de manera eficient. Aquest és un problema que afecta especialment els professionals intensius en informació en general i els que treballen en el món acadèmic en concret.

Aquest és l'àmbit en què seran més útils les generacions de sistemes del tipus PIM en línia, capaços d'integrar informació de diferents fonts, emmagatzemar-la i tenir-la sempre disponible per als usuaris a través de qualsevol ordinador, per mitjà del Web, o fins i tot en dispositius mòbils. Ja tenim una bona col·lecció d'aquestes aplicacions en línia, com RefWorks, 2collab o Connotea, però encara no han desplegat tot el potencial, com ho demostra la relativament escassa implantació en el món acadèmic. Això no obstant, les funcionalitats que ara ja presenten, com la possibilitat d'importar informació de fonts heterogènies i integrar-les en un sistema unificat i d'estar disponibles des de qualsevol lloc on hi hagi un ordinador i una connexió a Internet, donen una idea de les possibilitats futures, sobretot a mesura que les promeses del Web 3.0, del Web social i del Web semàntic es vagin fent realitat. Tenim davant nostre uns anys interessants.

4. Referències

Casárez, Vince [*et al.*] (2009). *Reshaping your business with Web 2.0: Using the new collaborative technologies to lead business transformation.* New York: McGraw Hill.

Codina, Lluís (2009). *Web 2.0 y Web 3.0 (diagrama interactivo)* [en línia]. <http://tinyurl.com/bzp57z> [Consulta: 29 maig 2009].

Codina, Lluís; Marcos, M. Carmen; Pedraza, Rafael (2009). *Web semántica y sistemas de información documental*. Gijón: Trea.

Díaz Noci, Javier [*et al.*] (2009). «Content and message analysis of online journalism: Some methodological proposals». *Trípodos*, núm. extra.

Feigenbaum, Lee [*et al.*]. «The Semantic Web in action». *Scientific American* (desembre).

Gobernor, James, Hinchcliffe, Dion; Nickull, Duane (2009). *Web 2.0 architectures*. Sebastopol: O'Reilly.

Gruber, Tom (2008). «Collective knowledge systems: Where the Social Web meets the Semantic Web». *Web Semantics: Science, Services and Agents on the World Wide Web* 6 (octubre), p. 4-13.

Lassila, Ora; Hendler, James. (2007). «Embracing Web 3.0». IEEE Internet Computing (maig-juny), p. 90-93.

Newitz, Annalee (2008). «Web 3.0. Playing it safe with our data». *The New Scientist*, fasc. 2647 (15 març), p. 42-43.

O'Reilly, Tim; Battelle, John (2009). *Web squared: Web 2.0 five years on* [en línia]. San Francisco: O'Reilly Media. <http://assets.en.oreilly.com/1/event/28/web2009_websquared-whitepaper.pdf> [Consulta: 29 maig 2009].

PEDRAZA JIMÉNEZ, Rafael; CODINA, Lluís; ROVIRA, Cristòfol (2008). «Semantic web adoption: Online tools for web evaluation and metadata extraction». A: RUAN, D. [*et al.*] (ed.). *Computational intelligence in decision and control. Proceedings of the 8th international FLINS conference*. New Jersey: World Scientific, p. 121-127.

— (2009). «Sistemas de información y metadatos en la web semántica». A: CODINA, Lluís; MARCOS, M. Carmen; PEDRAZA, Rafael (ed.). *Web semántica y sistemas de información documental*. Gijón: Trea, p. 1-42.

PORTER, Joshua (2009). *Designing for the Social Web*. Berkeley: New Riders.

RODRÍGUEZ MARTÍNEZ, Ruth; PEDRAZA JIMÉNEZ, Rafael (2009). *Hipertext.Net* [en línia], vol. 7 (maig). <http://www.hipertext.net/web/pag297.htm> [Consulta: 29 maig 2009].

SHIS, Clara (2009). *The Facebook era.* Boston: Prentice Hall.

SHUEN, Amy (2008). *Web 2.0: A strategy guide*. Sebastopol: O'Reilly.

Sessió II

Comunicacions

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2019, p. 119-124 DOI: 10.2436/15.2503.02.7

Vocabulària: un multicercador temàtic

Xavier Albons, Pep Cara, Àngels Egea, Montserrat Lleopart
Serveis Lingüístics
Universitat de Barcelona

Resum

Aquesta comunicació presenta el projecte Vocabulària, desenvolupat per la Universitat de Barcelona, que pretén la difusió de la terminologia catalana correcta. S'exposen els antecedents del projecte i els multicercadors que utilitza, i s'explica la gestió que fa dels continguts i la presentació dels resultats.

Paraules clau: bloc, multicercador, terminologia, Vocabulària.

Abstract: *Vocabulària: a theme-based multi-search engine*

This communication presents the Vocabulària project developed by the University of Barcelona, which seeks to disseminate correct Catalan terminology. The background of the project is outlined, and the multi-search engines it uses, and the way it manages contents and presents results is explained.

Key words: blog, multi-search engine, terminology, Vocabulària.

1. Introducció

Vocabulària és un projecte desenvolupat pels Serveis Lingüístics de la Universitat de Barcelona (UB) (*http://www.ub.edu/sl*) en què han participat el professorat i l'alumnat de la Xarxa de Dinamització Lingüística. Té l'objectiu de facilitar l'accés a la terminologia correcta usada en cada àmbit de coneixement mitjançant l'ús de multicercadors i la difusió d'obres en línia.

2. Antecedents del Vocabulària

Fa una vintena d'anys, alumnes de la Facultat de Química van fer un recull inicial de termes de química a partir del buidatge dels seus apunts de classe. Aquest

material de partida, amb l'aportació conceptual del professorat i metodològica dels Serveis Lingüístics (aleshores Servei de Llengua Catalana), va ser l'embrió de la col·lecció de «Vocabularis Bàsics per a l'Alumnat» (*www.ub.cat/enllaca/directori.php?branca=498*). Els vocabularis d'aquesta col·lecció, uns llibrets que recullen la terminologia bàsica de les matèries tractades a la UB amb equivalències en castellà i en anglès principalment, han acomplert durant gairebé dues dècades l'objectiu inicial de difusió de la terminologia catalana correcta, i més endavant han esdevingut un recurs de suport per a alumnat nouvingut i professorat visitant.

Els avenços tecnològics i la implantació de les noves tecnologies de la informació i la comunicació dins i fora de la comunitat universitària han permès, d'una banda, passar del format en paper al format electrònic —més barat de produir i més fàcil de difondre i d'actualitzar— i, de l'altra, han millorat les possibilitats de cerca d'informació i, per tant, d'aprofitament d'altres recursos en línia. Això és especialment important si tenim en compte la magnitud de la comunitat universitària: uns 81.000 estudiants (comptant-hi els de formació continuada), 4.700 professors i 2.200 treballadors d'administració i serveis, i la renovació periòdica d'una bona part d'aquest col·lectiu.

En aquest marc neix el projecte Vocabulària, que manté els mateixos objectius que la col·lecció de «Vocabularis Bàsics per a l'Alumnat» (difusió de la terminologia catalana correcta), però que, aprofitant les possibilitats que ofereixen les noves tecnologies, vol difondre la terminologia d'elaboració pròpia i també la de totes les obres consultables en línia en un sol corpus de consulta global o temàtica. La tecnologia de base per al desenvolupament del Vocabulària són els multicercadors.

3. Els multicercadors

Els multicercadors dels Serveis Lingüístics de la UB estan inspirats en el OneLook (*www.onelook.com*), un motor de cerca que conté, indexats i classificats temàticament, més de mil diccionaris en anglès, en els quals es poden fer cerques simultànies amb una resposta immediata i exhaustiva en què es detallen totes les fonts que contenen la cadena cercada i l'àrea temàtica a què pertanyen.

És una eina de cerca d'informació ràpida, perquè es fa una sola cerca a tot el conjunt i la resposta és immediata perquè la informació està indexada; i és una eina exhaustiva perquè inclou un nombre elevat de recursos d'una tipologia determinada. A més, presenta l'avantatge que l'usuari no necessita conèixer els recursos que li poden ser útils per trobar una determinada informació, sinó que el mateix motor de cerca mostra en la resposta els recursos que contenen la informació cercada. Així, un multicercador d'aquestes característiques és útil també com a eina de difusió de recursos.

Des de la generalització d'Internet com a mitjà d'informació i comunicació, la presència de recursos lexicogràfics i terminològics en català ha estat molt important. La major part de les institucions productores d'aquests recursos han optat per difondre'ls en línia: el *Diccionari de la llengua catalana* (DIEC) de l'Institut d'Estudis Catalans (IEC); *L'enciclopèdia*, el *Gran diccionari de la llengua catalana* (GDLC) i el *Diccionari enciclopèdic de medicina* (DEM) (Grup Enciclopèdia Catalana); Cercaterm (TERMCAT), etcètera.

La importància d'Internet per a la difusió en línia ha estat decisiva en el cas dels recursos lexicogràfics més petits, com ara els vocabularis de la col·lecció «Vocabularis Bàsics per a l'Alumnat» de la UB, i moltes altres col·leccions similars elaborades en altres universitats catalanes o les obres terminològiques elaborades des de molts altres organismes, com ara la Generalitat de Catalunya o el Consorci per a la Normalització Lingüística (CPNL).

Hi ha hagut iniciatives, fins i tot institucionals —com ara la de Llengua.org (*www.llengua.org*)—, per a fer directoris exhaustius de recursos en línia sobre diversos aspectes de la llengua catalana. Només cal fer-hi un cop d'ull per a adonarnos de la gran quantitat de recursos existents i, per tant, de la dificultat de conèixer-los tots i poder-los consultar.

Així, doncs, la gran producció d'obres terminològiques disponibles en línia s'ha acabat convertint en un inconvenient per a arribar als possibles usuaris. En la pràctica, els professionals de la llengua i altres usuaris tendeixen a consultar únicament els grans recursos en línia: DIEC (*dlc.iec.cat*), GDLC (*www.enciclopedia.cat*), *L'enciclopèdia* (*www.enciclopedia.cat*), *Cercaterm* (*www.termcat.cat*), *Diccionari enciclopèdic de medicina* (*www.grec.net/home/cel/mdicc.htm*), etc. I la resta de recursos, de dimensions menors, són ignorats i, per tant, desaprofitats, tot i que la suma de tots plegats pot arribar a representar un volum de dades igual o superior a alguns dels grans recursos.

En aquest context, els Serveis Lingüístics de la UB es van plantejar la necessitat de desenvolupar un multicercador terminològic que tingués unes prestacions similars al OneLook, és a dir, una eina que permetés tenir indexats tots els recursos existents d'una determinada tipologia per a poder-hi fer cerques simultàniament.

Es va observar que pràcticament tots els recursos susceptibles de ser integrats en el multicercador terminològic estaven indexats pels motors del Google, és a dir, eren consultables des del cercador Google. Per tant, si podíem aprofitar aquesta plataforma per a construir el nostre multicercador, no hi havia necessitat de construir un corpus de recursos indexats *ad hoc* —com es fa en l'Optimot (*optimot.gencat.cat*)—, amb totes les dificultats tècniques que això podia comportar i les dificultats d'actualització dels continguts quan la font originària s'ha modificat. A més, la indexació garanteix la rapidesa en la resposta, que no està garantida si no

hi ha indexació prèvia, com passa amb el Metacercador de la UPC (*www.upc.es/slt/metacercador*).

Només calia trobar la manera de restringir l'univers de cerca al conjunt de recursos que es volia incloure en el multicercador. El mateix Google ens va oferir la manera de construir el multicercador terminològic gràcies a la possibilitat de construir motors de cerca personalitzats (*www.google.com/coop/cse*). La construcció d'un motor de cerca personalitzat consisteix, simplement, en la introducció dels localitzadors universals de recursos (URL) dels recursos seleccionats en el motor de cerca.

La part més laboriosa de l'elaboració del multicercador va ser la confecció de la llista exhaustiva de tots els recursos que havia de contenir i el manteniment posterior dels URL d'aquests recursos, que es fa trimestralment.

Atesa la naturalesa dels recursos seleccionats, i amb l'objectiu de millorar els resultats de la cerca, es va considerar preferible de construir dos motors de cerca o multicercadors: un de terminològic i un de lingüístic.

El projecte Vocabulària parteix del multicercador terminològic i en fa un desenvolupament pensat per oferir a cada perfil d'usuari els recursos que poden ser més útils.

4. Implementació i descripció del Vocabulària

El Vocabulària és una interfície de consultes terminològiques temàtiques i també una plataforma d'informació i difusió feta amb WordPress, un sistema de gestió de continguts de codi obert. Aquest sistema és un dels més usats per a la publicació de blocs a la xarxa. Els gestors de continguts, com ara WordPress, són sistemes que permeten fer i actualitzar webs amb molta facilitat.

Els blocs (figura 1) tenen una pàgina principal o dinàmica on es van publicant notícies o entrades classificades o ordenades per categories o etiquetes temàtiques i, a part, poden tenir diverses pàgines estàtiques. Tant la pàgina dinàmica com les estàtiques comparteixen les barres laterals, la capçalera i el peu de pàgina.

A la pàgina dinàmica hi ha les informacions que són notícia, cadascuna amb un descriptor que les identifica per categories. Aquestes categories es corresponen temàticament amb les facultats —soles o agrupades temàticament— de la Universitat de Barcelona, per exemple, *biologia*, *química*, *ciències de la salut*, *dret*, *economia*, *educació*, *física*, *geologia*, *humanitats*, *matemàtiques*, i, finalment, també una categoria *general* per a les notícies que no corresponen temàticament a cap facultat o grup de facultats. En aquesta part dinàmica, es publicaran les novetats terminològiques fruit d'edicions o de consultes i altres notícies d'actualitat terminològica.

Pel que fa a les barres laterals, es va optar per un esquema estàtic o *tema* que té dues barres laterals. A la barra de l'esquerra hi ha enllaços a una sèrie de recur-

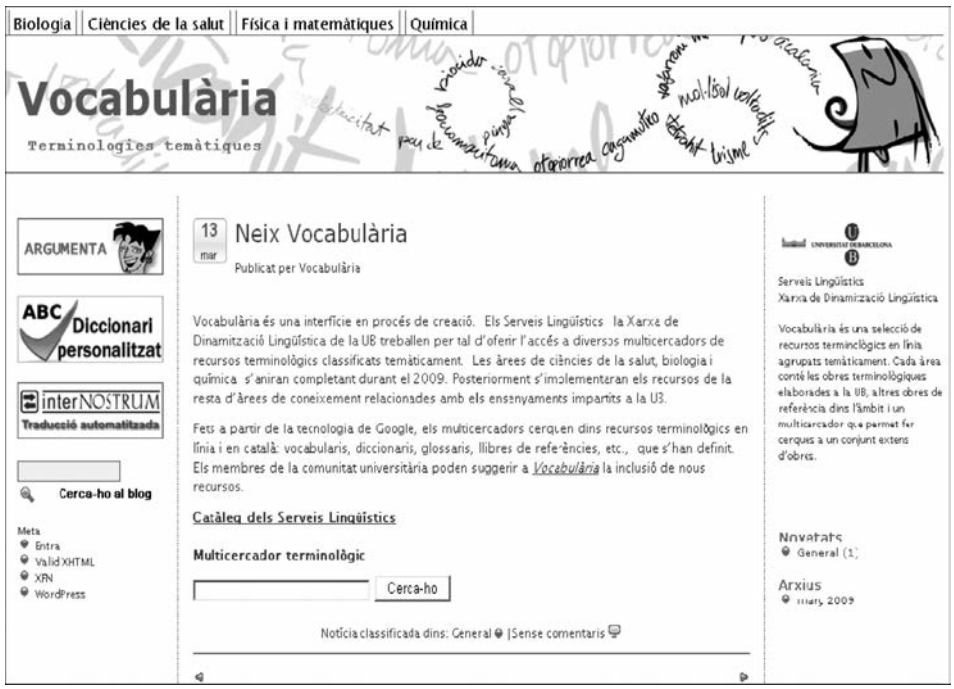


FIGURA 1. Portal del lloc web del projecte Vocabulària (*http://www2.ub.edu/sl/vocabularia*)

sos que estan en estreta relació amb la terminologia. S'ha enllaçat el gestor de consultes lingüístiques i terminològiques Sens Dubte, el diccionari personalitzat per als verificadors ortogràfics, el sistema de traducció automatitzada Internostrum i, finalment, la pàgina de recursos lingüístics de l'Àrea d'Assessorament Lingüístic i Terminologia dels Serveis Lingüístics de la UB, que conté altres eines per a cercar informació lingüística i terminològica o per a la redacció i edició de textos. A la barra de la dreta hi ha una breu descripció de què és el Vocabulària, qui el fa i també els arxius de les notícies de la pàgina principal ordenades per mesos i per categories.

Les pàgines estàtiques tenen una importància bàsica al web i s'hi pot accedir per unes pestanyes, molt visibles, situades sobre la capçalera. De fet, les diferents facultats poden fer un enllaç a la pàgina estàtica que els correspon, de manera que per a elles serà la pàgina principal del Vocabulària. Cada pàgina estàtica correspon a una facultat o un grup de facultats agrupades temàticament. Dins de cada pàgina estàtica hi ha:

1) El multicercador de la facultat o les facultats.
2) Els vocabularis en PDF elaborats pels Serveis Lingüístics i les comissions

de dinamització lingüística de cada centre, ubicats al dipòsit digital amb el seu identificador (*handle*) corresponent i on també cerca el multicercador.

3) Si es dóna el cas, altres recursos als quals no pot accedir el multicercador, com ara el *Diccionari enciclopèdic de medicina*.

De moment, s'han implementat les pàgines estàtiques corresponents a *biologia*, *química* i *ciències de la salut*, i progressivament s'aniran afegint la resta de facultats de la UB.

La participació del professorat i l'alumnat en el Vocabulària consisteix a assessorar sobre les obres terminològiques que formen el corpus de cada branca de cerca. Les comissions de dinamització lingüística de les facultats de la UB, on hi ha un representant de cada departament, donen el vistiplau a la selecció feta pels Serveis Lingüístics i poden fer propostes d'inclusió quan apareguin noves fonts interessants en el futur.

5. Desenvolupament futur

Els multicercadors presenten els resultats de cerca en l'ordre que Google té establert per defecte, que no sempre es correspon amb l'ordre de prioritat que es voldria donar a les fonts que constitueixen els multicercadors. En un futur proper està previst d'intervenir en la presentació dels resultats, de manera que apareguin en l'ordre que es consideri més convenient.

Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 125-133 DOI: 10.2436/15.2503.02.8

Terminologia i documentació 2.0

Jordi Chumillas, Ruth S. Contreras, Ricard Giramé
Universitat de Vic

Resum

La visió inicial del Web, el Web 1.0, es basava en pàgines estàtiques i sense interacció amb els usuaris. En canvi, el Web 2.0 és una evolució del Web que permet la publicació lliure d'informació, la reelaboració de continguts, la interacció dels usuaris i la creació de xarxes socials en evolució constant. La flexibilitat d'ús de les eines 2.0 és molt útil per als periodistes, els traductors, els docents, els estudiants, etc., perquè els facilita la cerca i la gestió de documents i de vocabulari especialitzat. En aquest estudi s'analitzen especialment tres d'aquestes eines: *AcronymFinder*, *SurveyMonkey* i *Forvo*.

Paraules clau: gestió de documents, vocabularis especialitzats, interacció persona-màquina, Web 1.0, Web 2.0, xarxes socials.

Abstract: *Terminology and documentation 2.0*

The initial view of the Web, the Web 1.0, was based on static pages with no user interaction. On the other hand, the Web 2.0 is an evolution of the Web that permits free information publishing, the re-doing of contents, user interaction and the creation of constantly evolving social networks. The flexibility of 2.0 tools is very useful for journalists, translators, teachers, students, etc., because they facilitate the search for and management of documents and specialised vocabulary. This communication particularly analyzes three of these tools: AcronymFinder, SurveyMonkey and Forvo.

Key words: document management, specialised vocabulary, people-machine interaction, Web 1.0, Web 2.0, social networks.

1. Introducció

Des que Tim Berners-Lee en va idear els principis i va contribuir a dissenyar-lo, el *World Wide Web* (WWW) s'ha convertit progressivament en un mitjà que

permet compartir amb d'altres usuaris experiències, activitats, oci o fins i tot feina. Moltes de les activitats quotidianes que portem a terme diàriament s'originen a la Xarxa o en depenen: el Web ens permet llegir el diari, operar amb el banc, fer la compra setmanal, demanar hora al metge, fer una partida de cartes, comunicar-nos amb els nostres amics i familiars o, fins i tot, fer la declaració de la renda. Des de fa uns quants anys, concretament des que el 2004 Tim O'Reilly va proposar el terme *Web 2.0*, el concepte de Web ha evolucionat i ha deixat enrere la visió inicial, basada en pàgines estàtiques que contenien diverses informacions, sovint poc actualitzades, i que poques vegades permetien la participació de l'usuari (això és, el Web 1.0).

Així doncs, el nou Web s'orienta cap a la interacció entre usuaris, la lliure publicació d'informació, la reelaboració constant de continguts i l'establiment d'autèntiques xarxes socials. Aquesta visió ha donat com a resultat l'aparició d'un bon nombre de pàgines i eines molt visuals i interactives (fins i tot n'hi ha que depenen exclusivament de la participació dels usuaris); unes pàgines que s'han convertit en un punt de trobada entre internautes de tot el món: plataformes com Facebook, Twitter o Wikipedia exemplifiquen a la perfecció aquesta nova realitat, que ha deixat de banda un sistema caduc basat en pàgines estàtiques i ha apostat pels usuaris, per les persones.

Precisament pel fet que, sovint, depèn directament dels usuaris, el Web 2.0 no és un recurs sistematitzat o organitzat d'una manera determinada, ni tampoc té una aplicació específica i inamovible: de fet, estableix xarxes obertes que evolucionen constantment, fins al punt que no és gens estrany que acabin perdent l'essència que tenien quan es van originar. Tornant a l'exemple de Facebook, tot i que Mark Zuckerberg el va idear com un punt de trobada fora de les aules entre alumnes de Harvard, actualment s'ha convertit en una eina de comunicació global, i algunes marques fins i tot l'empren per generar campanyes de fidelització.

La flexibilitat d'ús que ofereixen les eines del Web 2.0 ha fet que molts professionals de múltiples àmbits les tinguin en compte a l'hora de portar a terme algunes de les tasques que desenvolupen. Així doncs, en aquest estudi ens proposem seleccionar, presentar i avaluar diverses aplicacions 2.0 que poden resultar útils a aquells professionals les tasques dels quals exigeixen, en un moment o altre, la gestió i l'ús de documentació i llenguatges d'especialitat (professionals del periodisme, de la traducció, de la comunicació en general, docents, estudiants, etc.). No es tracta d'aplicacions ideades estrictament i exclusiva per a la pràctica de la terminologia/terminografia o la documentació, però sí que tenen aplicacions evidents en aquests camps i poden arribar a millorar-ne i facilitar-ne algunes de les tasques més quotidianes: cerca i gestió documental, cerca i gestió de vocabulari especialitzat, difusió del coneixement, etc. Tots aquests llocs web s'han triat per la

funcionalitat que tenen en l'àmbit d'estudi d'aquesta Jornada i s'han avaluat a partir de diversos paràmetres i indicadors per a l'anàlisi i l'avaluació de recursos digitals en línia.

2. Metodologia

En primer lloc, considerem important esmentar que ens enfrontem a una realitat nova i canviant i que, per tant, no té una tradició de recerca consolidada.

Per començar, hem reunit un corpus d'eines i aplicacions que podrien ser útils en les tasques de gestió documental i de llenguatges d'especialitat.

Seguidament, hem tingut en compte els següents principis de Tim O'Reilly a l'hora de filtrar quines de les eines es poden considerar Web 2.0 i quines no:

— La Xarxa n'és la plataforma.
— Aprofiten la intel·ligència col·lectiva.
— Allò que fa moure Internet és la informació.
— Allò que s'ofereix és un servei, no pas un producte que s'ha d'actualitzar.
— Tenen un model de programació lleugera que prioritza la simplicitat.
— Ofereixen serveis disponibles en qualsevol plataforma (PC, Mac, PDA, telèfon mòbil...).
— No es limiten a oferir continguts, sinó que ofereixen una experiència a l'usuari.

Tot i que hi ha eines que no segueixen tots aquests principis, hem cregut adient tenir-les en compte en l'estudi, sigui pel seu disseny, contingut o funcionalitat. El resultat final d'aquesta tria ha estat un llistat d'una trentena d'eines, de les quals n'hem avaluat nou (per cenyir-nos a l'espai disponible, però, en presentarem només tres):

— Scribd, permet publicar i compartir documents originals.
— Ebiwrite, eines per a la pràctica de la traducció.
— BackupURL, permet fer i gestionar còpies de seguretat de pàgines web.
— Mindomo, permet dissenyar arbres conceptuals.
— *Forvo*, diccionari multilingüe de pronúncia.
— *Lingoz*, diccionari multilingüe col·laboratiu.
— *Acronymfinder*, diccionari de sigles i acrònims.
— GoogleScholar, permet buscar bibliografia especialitzada.
— SurveyMonkey, permet crear i publicar enquestes.

Tenint en compte la diversitat d'eines i el fet que s'adrecen a usuaris diferents, es fa molt difícil fer-ne una comparació. Per aquest motiu, l'anàlisi que en proposem es basa en la funcionalitat de les eines i té en compte l'aplicació que se'n fa en

activitats de gestió documental i terminològica, en fa una descripció/definició, n'avalua els punts forts i febles i les classifica segons aquesta estructura que proposen Cobo i Pardo (2007), del Grup de Recerca d'Interaccions Digitals de la Universitat de Vic (UVic):

— Les xarxes socials permeten crear espais que promouen o faciliten la conformació de comunitats i instàncies d'intercanvi social.

— Els continguts afavoreixen la lectura, escriptura, distribució i intercanvi en línia.

— L'organització de la informació és social i intel·ligent. Presenta recursos per etiquetar, sindicar i indexar; i a més faciliten la classificació, l'ordenació i el dipòsit de la informació.

— Aplicacions i serveis (*mashups*), que són eines, aplicacions, plataformes en línia i híbrids de recursos creats per a oferir serveis de valor afegit a l'usuari final.

Per a la descripció i avaluació de les eines, hem tingut en compte els aspectes d'usabilitat següents:

— Opcions d'interacció i participació; és a dir, el paper de l'usuari en la creació i edició de continguts i opcions de comunicació amb d'altres usuaris.

— Navegació i recuperació; és a dir, l'estructura i l'accés a la informació és eficient en l'ús i la consulta.

El concepte d'usabilitat que hem introduït abans, el podríem entendre com la mesura en què un producte pot ser utilitzat per usuaris definits per a assolir objectius específics amb efectivitat, eficiència i satisfacció en un context d'ús concret. Nielsen (1999-2009) exposa en aquest context que la utilitat d'un sistema, entès com un mitjà per a assolir un objectiu, ha de posseir un component de funcionalitat (utilitat funcional) i un altre de basat en la forma en què els usuaris poden fer servir aquesta funcionalitat.

3. Avaluació de tres eines representatives

Arribats a aquest punt, ens disposem a presentar les tres eines que, segons el nostre criteri, resulten més representatives de la tasca feta:

— *Acronym Finder* (*http://www.acronymfinder.com*),

— SurveyMonkey (*http://www.surveymonkey.com*),

— *Forvo* (*http://www.forvo.com*).

3.1. *Acronym Finder*

3.1.1. Identificació de l'eina

Acronym Finder és un diccionari d'acrònims, símbols i sigles. Conté entrades en múltiples idiomes, tot i que l'anglès n'és la llengua vehicular. Es podria classificar com una eina basada en els continguts, destinada principalment a usuaris especialitzats (professionals de la docència, la traducció i la interpretació, el periodisme, la terminologia, estudiants universitaris, etc.). Ofereix el desenvolupament en llengua original de prop de 4 milions de sigles, acrònims i símbols, així com l'àmbit d'especialitat en què s'empren i, si escau, l'enllaç cap a la definició que se'n fa a *The Free Dictionary, by Farlex*. Com a exemple, mostrarem les entrades corresponents a *IEC* i a *SCATERM*.

3.1.2. Opcions destacades d'interacció i participació

Aquesta eina ofereix diverses opcions de participació a l'usuari. En primer lloc, permet compartir i publicar els resultats de cerca obtinguts mitjançant enllaços directes a diverses xarxes socials (Facebook, Delicious, Stumble Upon, Digg.com, Reddit i d'altres).

Permet, mitjançant un formulari en línia, suggerir als mantenidors de la pàgina la incorporació de nous acrònims. Els editors d'*Acronym Finder* avaluaran les noves propostes abans de donar-les per bones. Precisament, mitjançant aquest formulari hem inclòs al repertori de la pàgina l'entrada *SCATERM*.

Els responsables de l'eina mantenen un bloc on publiquen entrades que informen de les novetats incorporades a *Acronym Finder* o de tot allò relacionat amb l'àmbit dels acrònims, símbols i sigles. Com en qualsevol bloc, els usuaris tenen l'opció de comentar totes les entrades que s'hi fan.

3.1.3. Punts forts de l'eina

— No cal registrar-s'hi per a emprar-la.

— S'especialitza en uns continguts que no sempre apareixen en d'altres repertoris més coneguts.

— Permet múltiples àmbits d'activitat.

— Permet col·laborar en el creixement del recull.

— Ofereix una cerca ràpida, senzilla i intuïtiva.

3.1.4. Punts febles de l'eina

— Té poques opcions de recuperació de la informació. Relacionat amb això, no permet fer cap mena de cerca avançada.

— Només té l'anglès com a llengua vehicular.

3.2. *SurveyMonkey*

3.2.1. Identificació de l'eina

SurveyMonkey permet dissenyar i gestionar enquestes en línia. Basa el seu funcionament en tres grans eixos, que són el disseny del qüestionari a través del navegador web, la recollida de respostes i, finalment, el seguiment de resultats mitjançant una pàgina web que s'actualitza automàticament i en temps real. Es podria classificar com una aplicació que ofereix serveis a l'usuari final. Des del punt de vista del tema d'aquesta Jornada, té aplicacions evidents en activitats com ara l'elaboració d'enquestes per a estudis d'implantació terminològica o per a l'avaluació de la satisfacció dels usuaris de serveis terminològics i de documentació. Per tal de poder mostrar el funcionament de l'eina, hi hem adaptat el qüestionari sobre vocabulari esportiu que Marina Nogué i Xavier Vila presenten en l'article «Entre el *hockey* i l'hoquei» publicat a *Estudis d'implantació terminològica* (Eumo i Termcat, 2007).

3.2.2. Opcions destacades d'interacció i participació

SurveyMonkey no es pot considerar una eina col·laborativa del tipus xarxa social (Facebook, per exemple) perquè, de fet, no permet (ni ho pretén) que diferents usuaris treballin conjuntament en l'elaboració d'uns mateixos continguts, això és, el disseny i preparació d'un mateix qüestionari. Les opcions d'interacció entre usuaris són limitades i gairebé sempre verticals: qui dissenya i elabora el qüestionari té l'opció de fer-lo arribar al públic escollit, que no cal que tingui un compte a SurveyMonkey; al seu torn, qui rep el qüestionari només té l'opció d'emplenar-lo i validar-lo perquè retorni a l'usuari emissor, que, si ho vol, té l'opció de compartir els resultats del seu qüestionari a la Xarxa. Finalment, l'eina no disposa de cap dispositiu de comunicació instantània.

3.2.3. Punts forts de l'eina

— Ofereix més d'una dotzena de models de pregunta que l'usuari pot adaptar a les seves necessitats.

— Ofereix la possibilitat de distribuir els qüestionaris mitjançant diversos canals: hiperenllaços que es poden enviar o inserir a una pàgina web, correu electrònic i finestres emergents.

— Filtra i tabula els resultats de cada qüestionari, i permet descarregar-ne un resum en diversos formats: CSV, XML, HTML I XSL.

3.2.4. Punts febles de l'eina

— Disposa d'unes opcions d'interacció molt limitades.

— Només té l'anglès com a llengua vehicular. Els qüestionaris, però, es poden redactar en qualsevol idioma.

3.3. *Forvo*

3.3.1. Identificació de l'eina

Forvo és un diccionari multilingüe de pronunciació. Conté més de 280.000 paraules en 217 llengües. Cada entrada inclou la pronúncia del mot corresponent i la geolocalització de l'usuari que l'ha enregistrada. Es podria classificar com una eina basada en els continguts, destinada a usuaris de tot tipus, tant especialistes com públic en general. Com a exemple, hem inclòs en el repertori les entrades corresponents a *Universitat de Vic* i *SCATERM*.

3.3.2. Opcions destacades d'interacció i participació

Forvo permet que l'usuari participi activament en l'elaboració dels continguts; relacionat amb això, permet enregistrar la pronúncia de qualsevol paraula i proposar-ne perquè les enregistrin d'altres usuaris. També ofereix l'opció de puntuar de l'1 al 5 la pronúncia d'altres usuaris i fer-ne comentaris. A més, la plataforma disposa d'un sistema de missatgeria intern que facilita la comunicació personal entre usuaris. Finalment, l'usuari pot fer un seguiment de l'activitat de *Forvo* a través de Twitter i Facebook.

3.3.3. Punts forts de l'eina

— No cal registrar-s'hi per a poder consultar les pronúncies que conté.

— Permet que l'usuari participi en l'establiment i el creixement dels continguts.

— Permet que l'usuari registrat faci un seguiment detallat tant de les paraules de nova agregació com de les que ha pronunciat ell mateix.

— Permet la descàrrega de qualsevol pronunciació en format MP3.
— Recull múltiples camps d'especialitat.
— Ofereix una cerca ràpida, senzilla i intuïtiva.
— Les llengües vehiculars són l'anglès i l'espanyol.

3.3.4. Punts febles de l'eina

— No s'ha de superar cap mena de filtre a l'hora d'agregar una altra pronunciació, i per això podem dir que no hi ha control sobre la qualitat final del producte.
— Té poques opcions de cerca.
— El sistema de traducció a l'espanyol o a l'anglès de les entrades està poc desenvolupat, ja que es fa automàticament mitjançant la plataforma Google Translate.

4. Conclusió

El Web 2.0 ofereix a l'usuari la possibilitat de reinterpretar els serveis que se li ofereixen. D'aquesta manera, moltes eines que han estat creades amb finalitats específiques poden evolucionar fins al punt de ser aplicables en àmbits que, a l'origen, no havien tingut en compte (en l'àmbit que ens ocupa, la terminologia i la documentació, per exemple).

La naturalesa canviant d'aquestes eines i serveis (concepte del beta perpetu) en dificulta l'estudi sistematitzat: allò que resulta adient d'avaluar en un moment concret pot no tenir rellevància en un futur a causa de l'evolució que ha seguit l'eina.

Moltes de les eines analitzades presenten problemes d'usabilitat en les interfícies i formes d'interacció. En les noves aplicacions, la interacció hi té un pes més important, però els usuaris no tenen a la seva disposició un model clar sobre com funcionen. Les novetats generen confusió, i això exigeix a l'usuari un procés d'aprenentatge que pot arribar a ser difícil.

Les eines que hem avaluat en aquest estudi són el resultat d'una tria personal, fins i tot casual; ara bé, hi ha moltes altres eines que podrien resultar rellevants. Vegem-ne unes quantes:

— Footnote, biblioteca de documents històrics.
— Scirus, eina per a la recerca científica a la xarxa.
— Humyo, disc dur virtual que permet publicar, compartir i gestionar arxius de tota mena.
— CompareMyFiles, permet comparar diverses versions d'un mateix document i en marca les diferències.

— MindMeister, permet dissenyar i compartir mapes conceptuals a la xarxa.
— *Lexipedia*, xarxa semàntica en línia.
— *Shahi*, diccionari visual en línia.
— Dimdim, eina que permet fer videoconferències en línia.
— Doodle, gestor d'enquestes en línia.

Finalment, el resultat de l'avaluació de les eines està subjecte a la nostra percepció personal. Per tal d'aconseguir resultats més representatius, queda obert per a futures recerques comprovar amb usuaris reals que l'ús que proposem per a aquestes eines realment resulta eficaç. Aquest altre estudi també permetria aventurar l'èxit de les eines en qüestió des del punt de vista de la satisfacció dels usuaris.

5. Bibliografia

Berners-Lee, Tim (2000). *Tejiendo la red: El inventor del Worl Wide Web nos descubre su origen*. Madrid: Siglo XX.

Cobo, Cristóbal; Pardo, Hugo (2007). *Planeta Web 2.0*: *Inteligencia Colectiva o medios fast food* [en línia]. Barcelona: Mèxic. <http://www.planetaweb2.net> [Consulta: 5 maig 2009].

Codina, Lluís (2000). *Evaluación de calidad en sitios web*. Barcelona: Universitat Pompeu Fabra.

Nielsen, Jacob (1996-2009). *Top Ten Design Mistakes* [en línia]. <http://www.useit.com/alertbox/9605.html> [Consulta: 5 maig 2009].

O'Reilly, Tim (2005). *What is web 2-0? Design Patterns and Business Models for the Next Generation of software* [en línia]. <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> [Consulta: 5 maig 2009]

Vila i Moreno, F. Xavier [*et al.*] (2007). *Estudis d'implantació terminològica*. Vic: Eumo: Termcat.

Balanç i conclusions de la VII Jornada de la SCATERM

Marina Salse, Jaume Martí
Coordinadors de la Jornada

Les ponències i les comunicacions que s'han presentat en aquesta Jornada s'orienten en dos sentits, que reflecteixen en bona part les tendències actuals del treball investigador en els camps en què s'ha centrat.

D'una banda, reflecteixen una preocupació considerable per la gran acumulació d'informació que significa Internet. La Xarxa esdevé un camp de treball, experimentació i investigació molt important que fa prendre formes noves a tècniques de treball usades d'antic:

— Indexació per llenguatges controlats. Usats tradicionalment, continuen ara estant presents a la xarxa, però actualment prenen la forma de tendències noves i de vegades parteixen d'un tractament automàtic de base de la informació. Així, com ens afirmava el doctor José Luis Alonso Berrocal, els tesaurus tradicionals apareixen ara no solament com una eina d'organització de la informació en un entorn no automatitzat, sinó també com a expansions de la consulta en entorns informàtics. A part d'això, el doctor Lluís Codina ens ha parlat de les noves formes de llenguatge controlat aplicades a Internet, com ara les taxonomies, les ontologies o les folksonomies. També podem considerar sistemes de classificació o llenguatges controlats els etiquetatges que es fan de determinats termes per facilitar l'estructuració i la consulta dels recursos electrònics. Un exemple d'això, l'han posat els membres dels Serveis Lingüístics de la Universitat de Barcelona amb llur Vocabulària. Tanmateix, encara les iniciatives són molt aïllades. Els grans projectes, com el Web semàntic, del qual ens ha parlat el doctor Codina, estan de moment poc madurs i no sabem si reeixiran.

— Indexació automàtica. Els estudis en aquest camp es van iniciar els anys seixanta i es van vincular a l'explotació de bases de dades. L'aparició d'Internet i dels seus milions i milions de pàgines ha fet que s'hagin hagut de buscar sistemes per a aconseguir cada vegada més bons resultats, atès que la indexació humana

hauria necessitat una gran inversió de temps i diners. Així, doncs, s'han fet i es fan nombrosos estudis per millorar la indexació automàtica i per crear resums automàtics. Les tendències en investigació han portat a crear dos grans tipus de mètodes: els no lingüístics, especialment centrats en l'estadística i el càlcul de probabilitats, i els lingüístics, que poden anar des de la simple lematització fins a l'ús de tesaurus i/o corpus documentals per a millorar el tractament terminològic de la indexació controlada i, en conseqüència, també la recuperació de la informació. Tot i que normalment els sistemes d'indexació solen usar una combinació d'ambdós tipus de mètodes, en la Jornada d'avui se'ns han presentat essencialment mètodes de tipus estadístic, com el model vectorial, mostrat pel doctor Alonso Berrocal o bé els estudis de Rogelio Nazar i dels professors de la Universitat Oberta de Catalunya, Mercè Vázquez i Antoni Oliver.

Ara hi ha, doncs, un important maridatge entre estadística, lingüística i informàtica. Lletres i ciències, abans tan separades en els plans d'estudis, s'han unit gràcies a la Xarxa i a les noves tendències d'investigació en indexació i recuperació de la informació.

D'altra banda, reflecteixen la vigència dels treballs clàssics de terminologia. El «Vocabulari de preservació i conservació del patrimoni documental» de la professora Maria Elvira de la UB és una d'aquestes tasques en què la informàtica només intervé com a eina auxiliar. És necessari que es continuïn fent treballs d'aquesta mena per a consolidar la terminologia en els diferents camps del coneixement i incorporar-hi els neologismes; treballs orientats, però, segons els objectius específics, entre els quals destaca el de constituir la base per a un millor tractament automatitzat de la informació i les bases de dades.

Pel balanç fet fins aquí, podem afirmar que les línies més evidents de continuïtat en la recerca pel que fa als temes tractats en aquesta Jornada són en el camp de la recuperació d'informació i en el desenvolupament i l'ús de les xarxes d'Internet, amb el maridatge que esmentàvem entre estadística, lingüística i informàtica.

I, quant a la terminologia, ens ha aparegut en aquesta Jornada com un fi en si mateix, pel valor que té a l'hora de facilitar la comunicació i de resoldre els problemes d'ús lingüístic que en deriven, en l'elaboració de diccionaris especialitzats i en el disseny i la utilització de multicercadors. Però, la presència de la terminologia, l'hem tinguda sobretot com a element auxiliar imprescindible en les operacions de cerca i recuperació d'informació, pròpies dels documentalistes i també dels traductors.

Per a totes aquestes funcions, la terminologia ha de consolidar conceptes i modes de tractament, per a afinar la resposta a qüestions que directament o indirecta han sorgit en aquesta Jornada: el tractament dels termes que també són lèxic comú i les decisions conseqüents sobre inclusió i contingut de les definicions; els

criteris per a la constitució de corpus especialitzats per a contrastar resultats estadístics amb els dels corpus de llengua comuna o general, amb fronteres que cal distingir de les fronteres temàtiques.

No falten, doncs, temes per a futures reflexions i debats en marcs acadèmics com el d'aquesta Jornada de «Terminologia i documentació».

Assistents a la VII Jornada

Aquesta llista recull totes les persones assistents a la VII Jornada de la SCATERM ordenades alfabèticament pels cognoms.

Xavier Albons Gomila
Barcelona

Salvador Alegret i Sanromà
Barcelona

José Luis Alonso Berrocal
Salamanca

Albert Amat
Barcelona

Elena Aragón Palancar
Cornellà de Llobregat

Sílvia Argudo Plans
Barcelona

Carme Bach Martorell
Barcelona

Marc Barracó i Serra
Barcelona

Miquel Centelles Velilla
Barcelona

Jordi Chumillas i Coromina
Vic

Lluís Codina Bonilla
Barcelona

Mireia Comas Via
Viladecans

Àngels Egea i Puigventós
Barcelona

Maria Elvira i Silleras
Barcelona

Agustí Espallargas i Majó
Barcelona

Constança Espelt Busquets
Barcelona

Anna Font Reñé
Barcelona

Gemma Fonrodona Baldajos
Barcelona

Glòria Fontova Hugas
Barcelona

Francesc Galera Porta
Sabadell

Mercè Gálvez Flaqué
Barcelona

Ricard Giramé Parareda
Vic

Eivor Jordà Mathiasen
Massarrojos

Iban Jordà Sánchez
Barcelona

Núria Jornet Benito
Vilanova i la Geltrú

Josep Maria Jovells Salvia
Golmés

Montserrat Lleopart Grau
Barcelona

Sílvia Llovera Duran
Barcelona

Mercè Lorente Casafont
Barcelona

Ruxandra Lungu
Barcelona

Heura Marçal Serra
Barcelona

Jaume Martí i Llobet
Barcelona

M. Rosa Mateu Martínez
Barcelona

Josep M. Mestres i Serra
Barcelona

Eulàlia Miret Raspall
Sant Pere de Ribes

M. Amor Montané March
Barcelona

Rogelio Nazar
Barcelona

Antoni Oliver González
Barcelona

M. Mar Palomo Delgado
Santa Coloma de Gramenet

Lourdes Pascual Gargallo
Castelló de la Plana

Mario Pérez-Montoro Gutiérrez
Barcelona

Conxa Planas Planas
Barcelona

Lluc Potrony Julià
Barcelona

Carme Prats
Barcelona

Mireia Ribera Turró
Barcelona

Anna Rubió Rodon
Barcelona

Aina Rusca Mestre
Barcelona

Marina Salse Rovira
Barcelona

Margarida Sanjaume i Navarro
Barcelona

M. Rosa Seguí Palou
Barcelona

Mariona Torra Ginesta
Barcelona

Cristóbal Urbano Salido
Barcelona

Mercè Vázquez i Garcia
Barcelona

Laura Vinuesa Baliu
Barcelona

Enkeleda Xhelo Çomo
Barcelona

Lluís de Yzaguirre Maura
Barcelona

AQUESTA OBRA S'HA ACABAT D'IMPRIMIR
A L'OBRADOR DE LIMPERGRAF, SL,
A BARBERÀ DEL VALLÈS,
EL DIA 18 DE FEBRER DE 2010