UN MODULO LINGÜISTICO PARA LA BUSQUEDA INFORMATICA DEL LENGUAJE NATURAL

Por BLANCA DE MENDIZABAL ALLENDE

Licenciada en Filología Inglesa. Filóloga de CINADDE.

y MIGUEL LOPEZ-MUÑIZ GOÑI

Magistrado, profesor de la Facultad de Informática. Presidente de la Asociación de Informática y Derecho.

I. Introducción

El lenguaje jurídico tiene su base en el lenguaje común, afirmándose por Olivecrona (1) que «nuestro lenguaje legal es básicamente una parte del lenguaje corriente». Es más, «el primitivo lenguaje jurídico no es un lenguaje riguroso, sino común», dice Capella (2).

Cuando se trata de utilizar el lenguaje natural para la recuperación de documentación jurídica, es decir, en un sistema de *full text* puro, en donde los elementos de recuperación de la información van a obtenerse directamente del texto legal, hemos de partir de la base de que nos encontramos con todos los elementos de ambigüedad y de imprecisión que tiene el propio lenguaje natural.

(1) K. OLIVECRONA: «Legal Language and Reality», en *Essays in honour of Roscoe Pound.* Indianápolis. Bobbs-Merrill. Co. 1962.

(2) JUAN RAMÓN CAPELLA: *El Derecho como lenguaje.* Ed. Ariel, Barcelona. 1968.

La búsqueda de información solamente a través de un operador de truncadura, aunque ésta sea múltiple (prefijos, sufijos y enmascaramiento de caracteres intermedios) no resuelve todos los problemas.

La selección de las palabras en su verdadero sentido, sin poder acudir por el momento a los estudios de estructuras profundas del lenguaje, ha de permitir al usuario la utilización de un módulo del lenguaje natural que le dé posibilidades de determinación de palabras en su verdadero sentido, pero con apertura a criterios ampliatorios cuando las necesidades de búsqueda así lo demanden.

Cuando se trata de búsqueda a través de descriptores ya sabemos que tenemos el *thesaurus* como elemento disponible. Pero cuando tratamos de emplear exclusivamente el lenguaje natural, ¿no podremos establecer unas determinadas condiciones para que el ordenador ayude a buscar exactamente lo que se busca?

De ahí la idea de confeccionar un «Módulo lingüístico» que permite una utilización del lenguaje natural de forma lo más práctica y al mismo tiempo extensiva posible.

Hemos utilizado la palabra «Módulo» por ser el elemento lingüístico adecuado para servir de comparación en determinadas búsquedas del lenguaje natural, de tal forma que aquellas sean congruentes.

Partimos de la base de que se trata de un estudio teórico, que todavía no ha sido posible ensayar en la máquina.

II. Palabras vacías

Lo primero que habrá que hacer es descargar el texto de todas las palabras no significativas, construyendo lo que se ha llamado un *thesaurus negativo* o antithesaurus, puesto que si el thesaurus es una relación de términos útiles sobre los que se va a buscar la información, la relación de palabras vacías supone precisamente que nunca van a tenerse en cuenta a efectos de búsqueda.

Estas palabras vacías son de dos tipos:

1. *Palabras vacías absolutas*

Son todas aquellas que pueden afirmarse no son útiles como elementos de recuperación en ninguna base de datos.

Se incluyen entre las palabras vacías las siguientes:

— Artículos.
— Contracciones.

— Preposiciones.
— Conjunciones.
— Adjetivos.
— Pronombres.
— Adverbios.
— Adverbios nominales.
— Verbos auxiliares HABER, SER, TENER y ESTAR.

Esto supone que hemos eliminado un total de 180 palabras vacías más 236 formas verbales auxiliares, lo que supone 416 palabras en total.

No obstante, ha de hacerse constar que determinadas palabras no pueden ser consideradas como vacías absolutas, por poder tener un significado específico debido a su homonimia, por ejemplo,

HABER, de importancia en contabilidad y economía,
SER
ERA
ESTADO
Etc.

En el cuadro adjunto se incluye una lista de posibles palabras vacías.

Una forma de poder establecer la diferencia entre una palabra vacía y su homónima significativa es la de establecer una tabla de términos dudosos, para que el ordenador detecte su aparición en el documento y se ofrezca al lingüista responsable del sistema el contexto donde aparece la palabra homónima. El lingüista deberá establecer la validez o no del término, dejándolo como está si no tiene valor significativo, o haciendo una marcación de la palabra, si tiene validez, marcación que puede consistir en cualquier símbolo admitido por el ordenador y que no sea significativo para el programa de recuperación.

2. *Palabras vacías relativas*

Son todas aquellas que, en relación con la base de datos, no tienen significación.

Se trata de descargar las posibilidades de búsqueda mediante la inclusión, como palabras vacías de ciertos verbos, adjetivos e incluso sustantivos, que no se consideran útiles a los efectos de búsqueda.

Una forma de trabajo para la determinación de estas palabras vacías relativas es la confección de un listado de palabras con su frecuencia, y estudiar las de uso más frecuente que serán, lógicamente,

las menos significativas. Pero de todas formas, la determinación de incluir una palabra en la relación de vacías es de alta responsabilidad, pues en alguna ocasión puede producir la pérdida de información.

III. Locuciones vacías

En el lenguaje natural existen muchísimas locuciones que tienen el carácter de vacías, en cuanto que no suponen más que formas de relación en el discurso y no son significativas.

Por ejemplo:

NO HA LUGAR
A MEDIDA QUE
A RESERVA DE
EN TODO CASO
A SABER
NO OBSTANTE
EL JUEGO DE LOS ARTICULOS
A MENOS QUE
GRACIAS A
EL PUNTO DE VISTA
SU RAZON DE SER
EN EL SENTIDO DE

En estas locuciones están presentes palabras que nunca pueden ser consideradas, por sí mismas, vacías, y al mismo tiempo pierden su carácter de significativas cuando se integran en dicha locución. Por ejemplo:

LUGAR DE IMPOSICION
MEDIDA DE SEGURIDAD
RESERVA LEGAL
CASO FORTUITO
DERECHO DE GRACIA
RAZON SOCIAL
Etc.

son locuciones que tienen palabras que figuran entre las vacías, y que sin embargo deben estimarse plenamente significativas.

La forma de poder establecer estas locuciones vacías es doble. De una parte, mediante un estudio previo, señalando las de uso más frecuente, incluso llegando a un estudio de estructuras del lenguaje bastante complicado. De otra parte, mediante listas KWIC, ir compro-

bando sobre los textos concretos cuáles son las frases vacías que se contienen en cada documento.

Para suprimir las locuciones vacías puede utilizarse un sistema similar al de las palabras vacías, cual es la de hacer un análisis manual, subrayando o uniendo las diferentes palabras por un signo no significativo informáticamente, o bien establecer un tratamiento informático para que de forma automática se eliminen dichas frases.

Teniendo en cuenta que los sistemas de tratamiento de texto completo tienen normalmente operadores de distancia, y especialmente el operador de adjunción ADJ, puede realizarse una lista de locuciones vacías y establecer el programa de que cuando aparezcan en una determinada secuencia las palabras que integran la locución, automáticamente se eliminen todas las palabras del fichero inverso de búsqueda.

IV. **Grupos significativos de palabras**

Se trata del tema contrario al que hemos visto en el apartado anterior. Muchas palabras pueden tener un significado en sí mismas, pero junto a otras pasan a representar un concepto perfectamente definido y que es altamente significativo. Por ejemplo:

ERROR MANIFIESTO
ABUSO DEL DERECHO
AGENTE DE CAMBIO Y BOLSA
BANCO DE ESPAÑA
LEY DE ENJUICIAMIENTO CRIMINAL

son todos sintagmas perfectamente válidos tal y como aparecen en los documentos.

Ahora bien, en muchas ocasiones estos sintagmas no tienen el carácter de forzosos, y a los que ya nos hemos referido con anterioridad, sino que sus componentes pueden estar separados, segregados o incluso enmascarados.

Por ejemplo, existe desplazamiento en los casos siguientes:

— No puede hacerlo más Banco que el de España.
— Existió un ERROR que puede considerase como MANIFIESTO.

Otras veces, se trata de términos equivalentes y no aparece el sintagma forzoso, como por ejemplo:

...en la Ley de procedimiento criminal
en vez de

...en la Ley de enjuiciamiento criminal.

En el primer caso, puede lograrse corregir la pérdida posible de información mediante los operadores de párrafo o de distancia, a los que haremos referencia en el tema dedicado a la búsqueda de información.

Mayor inconveniente tiene el segundo aspecto, puesto que al tratarse de un sistema de búsqueda literal y no conceptual, el ordenador no puede determinar identidades de caracteres, lo que obligará al consultante a utilizar términos o palabras alternativas para no perder información.

V. Posible estructura de un módulo lingüístico

Entendemos por módulo lingüístico un programa que tiende al tratamiento del lenguaje natural de una forma estructurada siguiendo determinados criterios de tal manera que se puedan recuperar las diferentes formas de los signos lingüísticos.

Sin pretender hacer un estudio exhaustivo, y a mero título de ejemplo, expondremos algunas ideas sobre este tema:

1. *Estructura del módulo*

El módulo estaría formado por unas tablas de correlación, accesibles mediante determinados documentos, con el fin de que el usuario pueda determinar el nivel de precisión o de extensión que le fuera más conveniente en razón del número de documentos encontrados, el grado de pertinencia o de exhaustividad que pretenda obtener de su consulta.

Todas estas relaciones tendrían que ser determinadas a priori mediante los oportunos estudios lingüísticos.

Las relaciones que se establecerían serían gramaticales, sintácticas y conceptuales, conforme a los criterios que a continuación se exponen.

2. *Sustitutos gramaticales* (G)

Son todas aquellas formas de las palabras, tales como el género, el número y tiempos verbales.

Habría de distinguirse entre dos grandes grupos de palabras:

A) *Las formas regulares,* en que se determinarán las siguientes:

M	MASCULINO
F	FEMENINO

S	SINGULAR
P	PLURAL
V	FORMA VERBAL

De esta manera todos los substantivos y los verbos quedarían estructurados mediante una raíz y sus correspondientes sufijos, tales como:

F	A
P	S
	AS
	ES
	ES
	OS
V	DOS O MENOS CARACTERES
	TRES O MAS CARACTERES

Con lo anterior, se determinarían la raíz y los sufijos dentro de las formas regulares.

B) *Las formas irregulares.*—Existen muchísimas palabras que varían su raíz o incluso los sufijos cuando cambian de género o de número. Por ejemplo:

a) Existen palabras distintas para el masculino y el femenino: PADRE-MADRE; TORO-VACA; YERNO-NUERA.

b) Existen terminaciones especiales para algunas palabras, y no se sigue la regla general de terminación en «A»:

BARON	BARONESA
PROFETA	PROFETISA
GALLO	GALLINA
JABALI	JABALINA

Esto es mucho más frecuente en los verbos llamados por lo mismo «irregulares», que deberá preverse toda la relación de formas verbales para acoplarlas al infinitivo.

3. *Sustituciones sintácticas* (S)

Se trata de poder establecer todas las palabras derivadas, partiendo de una determinada raíz, sirviendo tanto las palabras simples como las compuestas.

En este caso, es muy importante poder determinar la raíz de los afijos, con el fin de determinar de manera automática en las palabras

todas las posibles conexiones de términos partiendo de una determinada raíz modificada por los prefijos o los sufijos.

A título de ejemplo citaremos los siguientes casos:

A) *Prefijos* (PR).—En la gramática se establece una serie de prefijos de diferente significación, tales como:

A	ASIMETRICO
ANTE	ANTEBRAZO
CO	COPROPIETARIO
CONTRA	CONTRAVENTANA
DES	DESHACER
EN	ENNEGRECER
ENTRE	ENTRESACAR
EXTRA	EXTRAORDINARIO
HIPER	HIPERTENSION
HIPO	HIPOTENSION
INFRA	INFRAUTILIZADO
INTER	INTERPONER
POS	POSPONER
RE	REEDIFICACION
SOBRE	SOBREPONER
SUB	SUBSUELO
SUPER	SUPERABUNDANCIA
Etc.	

B) *Sufijos de los sustantivos.*—Pueden ser de diferentes clases, tales como:

a) *Cualidad* (SC)

ANCIA	REPUGNANCIA
ENCIA	DEMENCIA
DAD	SUAVIDAD
EZA	CRUDEZA

b) *Substantivaciones verbales* (SU)

ANZA	COBRANZA
DOR	GANADOR
DURA	TORCEDURA
CION	RECLAMACION
MIENTO	PENSAMIENTO
ARIO	ARRENDATARIO

c) *Postverbales* (SP), tales como:

DE COSTAR	COSTE
DE PAGAR	PAGA
DE EMBARCAR	EMBARQUE
DE ABONAR	ABONO

d) *Aumentativos y diminutivos* (SA-SD), tales como:

ON	HOMBRON
AZO	PERRAZO
OTE	MUCHACHOTE
ITO	ARBOLITO
ILLO	CASILLA
Etc.	

e) *De profesión u oficio* (SD):

ANTE	CAMBIANTE
ARIO	BIBLIOTECARIO
DOR	BORRADOR
ERO	CAJERO
ISTA	PERIODISTA (sl)

f) *Colectivos:*

AL	ARENAL
EDO	ROSALEDA
AMEN	VELAMEN
AR	ATOCHAR

g) *Despectivos* (SE):

ACO	PAJARRACO
AJO	COLGAJO
UCO	JUMERUCA

C) *Sufijos de adjetivos* (A)

a) *De cualidad* (AC):

ADO	COLORADO
DERO	CRECEDERO
IENTO	HAMBRIENTO
IZO	ENFERMIZO
OSO	GRACIOSO

b) *Gentilicios:*

ANO	ASTURIANO
ENSE	GERUNDENSE
EÑO	MADRILEÑO

ES	FRANCES
INO	SALMANTINO
I	MARROQUI

c) *Aumentativos, despectivos, diminutivos* (AA, AD, AM):

ON	CABEZON
ACHO	RICACHO
ITO	GUAPITO
CITO	POBRECITO
Etc.	

4. *Sustituciones conceptuales* (C)

Son aquellas relaciones entre palabras que pueden permitir una búsqueda en cierta forma similar a la que puede establecerse por una relación de sinonimia en un thesaurus estructurado.

Esta relación conceptual únicamente puede establecerse a nivel humano. Por ejemplo:

VERIFICACION	CONTROL
CONTROL	DOMINIO
DOMINIO	PROPIEDAD
SUJETO PASIVO	SUJETO IMPONIBLE
PERJUICIO	DAÑO
EJERCICIO	AÑO
AÑO	AÑO NATURAL
COMPRADOR	ADQUIRENTE

Ahora bien, de esta relación vemos que existen algunas sustituciones que son válidas de forma directa, pero que no permiten establecer un automatismo en todas sus relaciones directas e inversas.

Las sustituciones conceptuales siguen determinadas reglas, que pueden enunciarse muy someramente de la siguiente forma:

a) Las sustituciones son reflexivas, es decir, que son válidas en un determinado orden pero pueden no serlo en el inverso.

CONTROL DOMINIO DEPENDENCIA

b) Las sustituciones son intransitivas, es decir, que son válidas en el orden en que se establecen dos entre sí, pero no son válidas entre las referenciales. Por ejemplo:

CONTROL	DOMINIO
DOMINIO	PROPIEDAD

Estas dos relaciones son válidas entre sí, pero no es válida establecer la sustitución:

CONTROL PROPIEDAD

Podríamos decir que si bien entre cada una de las primeras relaciones existe una sinonimia, entre sí no son términos absolutamente sinónimos, sino quasi-sinónimos.

c) Las sustituciones son antisimétricas, lo que es lo mismo a afirmar que la relación es válida en un orden pero no lo es en el contrario.

P1	P2
EJERCITO	AÑO

es válido, pero no siempre es válido

P2	P1
AÑO	EJERCITO

ya que únicamente es cierto cuando

P1 = P2

IV. Utilización del módulo lingüístico

Una vez establecidas todas las anteriores relaciones, lo cual no deja de ser muy complicado, el uso sería relativamente fácil. Bastaría con utilizar cualquiera de los comandos establecidos para recuperar determinada palabra o palabras.

1. *Utilización de la estructura gramatical* (G)

Como ya hemos indicado, podría accederse a través de los operadores.

F	FEMENINO
P	PLURAL
V	VERBAL
$	TRUNCADURA
$N	TRUNCADURA CON DISTANCIA

PREGUNTA COMPRADOR FP	COMPRADORA
	COMPRADORES
	COMPRADORAS

COMPRAR V	COMPRO	COMPRAS
	COMPRABA	COMPRADO
	COMPRARAN	Etc.
COMPRAS	COMPRADOR	
	COMPRADORA	
	COMPRADORES	
	COMPRADORAS	
	COMPRAR	
	COMPRADO	
	Etc.	
COMPRA	COMPRADOR	
	COMPRABA	
	COMPRADO	
	Etc.	

2. *Utilización de la estructura sintética* (S)

Como en el caso anterior, tendríamos los siguientes operadores:

PR	PREFIJOS
SU	SUFIJOS DE SUBSTANTIVOS
SC	SUFIJO CUALIDAD
SV	SUFIJO VERBAL
SP	SUFIJO POSTVERBAL
SA	SUFIJO AUMENTATIVO
SD	SUFIJO DIMINUTIVO
SO	SUFIJO OFICIO
SL	SUFIJO COLECTIVO
SE	SUFIJO DESPECTIVO
AD	SUFIJO DE ADJETIVOS
AC	SUFIJO DE CUALIDAD
AG	SUFIJO GENTILICIO
AA	SUFIJO AUMENTATIVO
AD	SUFIJO DESPECTIVO
AM	SUFIJO DIMINUTIVO
Etc.	

La pregunta podría formularse de la siguiente manera:

PREGUNTA	OPERADOR	CONTESTACION
ARRENDAR	SV	ARRENDADOR
		ARRENDATARIO
		ARRENDAMIENTO
DOMINIO	PR	CONDOMINIO

CONFORMAR	SC	CONFORMIDAD
	PR	DISCONFORMAR
	PR-SC	DISCONFORMIDAD

3. *Utilización de las dos estructuras* (S y G)

Como es natural, con ambas estructuras funcionando podría encontrarse todas las palabras cuya raíz fuera la buscada, y en todas sus manifestaciones: sustantivos simples o compuestos, en cualquier género y número, sustantivaciones, etc.:

ARRENDAR	S AND	G	ARRENDADOR
			ARRENDADORA
			ARRENDADORES
			ARRENDADORAS
			ARRENDAMIENTO
			ARRENDATARIO
			ARRENDATARIA
			ARRENDATARIAS
			ARRENDAMIENTO
			ARRENDAMIENTOS
			ARRIENDO
			ARRENDADO
			COARRENDAMIENTO
			COARRENDATARIO
			ARRENDABA
			ARRENDO
			Etc.

4. *Utilización de la estructura conceptual* (C)

Permitiría buscar cualquier palabra cuya sustitución conceptual hubiera sido determinada.

DOMINIO	C	PROPIEDAD

5. *Utilización conjunta de las estructuras conceptual y gramatical*

Avanzando en la complejidad de la pregunta, la respuesta podría ser la siguiente:

DOMINIO	CG	DOMINIO
		PROPIEDAD
		DUEÑO

DUEÑA
PROPIETARIO
PROPIETARIA
DUEÑOS
DUEÑAS
PROPIETARIOS
PROPIETARIAS
Etc.

6. *Utilización conjunta de las estructuras conceptual, gramatical y sintáctica* (C, G y S)

Es la última fórmula de búsqueda alternativa de palabras, y permite la recuperación de una serie de términos relacionados de múltiples maneras:

ADQUIRIR	CGS	ADQUIRENTE
		ADQUIRENTES
		ADQUIRIDO
		COMPRAR
		COMPRADOR
		COMPRADORA
		COMPRADORES
		COMPRADORAS
		COMPRADO
		COADQUIRENTE
		ADQUISICION
		COMPRAVENTA
		Etc.

SUMARIO

El Derecho se expresa con palabras del lenguaje natural; por ello, cuando se trata de buscar en bases de datos jurídicos a través del texto completo, es necesario utilizar sistemas que permitan una mayor precisión. Un módulo lingüístico como el que se ofrece en esta comunicación puede ser útil para el tratamiento del lenguaje natural en los ordenadores.