El Corpus Textual Informatitzat de la Llengua Catalana

Joan Soler i Bou

Institut d'Estudis Catalans

1. Antecedentes

La evolución experimentada por los procedimientos de constitución de corpus en los últimos años ha supuesto cambios notables en las posibilidades de planificación y ejecución de los proyectos dedicados a la creación de este tipo de recursos lingüísticos. En este contexto, el "Corpus Textual Informatitzat de la Llengua Catalana" (CTILC), por los años en que fue diseñado y constituido, puede ser tomado como un ejemplo previo a la generalización y sistematización de los procesos de constitución de corpus. El CTILC, además, es un corpus que fue terminado en 1997: se construyó en el ámbito de un proyecto lexicográfico de largo alcance, denominado "Diccionari del Català Contemporani" (DCC), que ya se encuentra en fase de redacción de un diccionario descriptivo a partir del corpus, por lo que en nuestro grupo de trabajo los aspectos de constitución de corpus no representan, desde el año 1998, una línea de trabajo efectiva.

De ahí que en mi intervención en estas jornadas mi objetivo no sea simplemente dar una descripción detallada de la estructura y los procesos de constitución del CTILC, sino propiciar, a través de su caracterización general, ciertas reflexiones: ¿como podemos, a partir de la experiencia adquirida en la constitución de corpus textuales, determinar qué aspectos son fundamentales en los corpus de referencia?, ¿en qué medida las nuevas posibilidades pueden condicionar el diseño y la (re)utilización de los corpus constituidos en la actualidad? ¿Como inciden esas posibilidades en la representatividad de los corpus de referencia? Esas son simplemente algunas de las preguntas que podemos plantearnos y parece claro que ninguna de ellas tiene una respuesta sencilla ni única. En todo caso, creo que vale la pena introducirlas en este contexto, que es precisamente el de las acciones previa de diseño y de planificación de un corpus de referencia para el Euskera del siglo XXI.

2. El proyecto DCC: esquema general

El CTILC es una parte fundamental del proyecto DCC, que desarrolla desde 1985 el Institut d'Estudis Catalans (IEC). La Sección Filológica del IEC desempeña las funciones de academia de la lengua catalana, por lo que estableció el proyecto DCC como una acción complementaria, en el ámbito de la constitución de recursos lingüísticos y de los trabajos descriptivos, para que el desarrollo de estas funciones pudiera hacerse con un mayor conocimiento empírico del estado actual de la lengua catalana contemporánea.

El desarrollo temporal del DCC se estructura en dos fases:

a) Fase de constitución de recursos lingüísticos: fundamentalmente, estos recursos son un corpus textual (el CTILC), que se utiliza como referencia descriptiva en los trabajos

de redacción del diccionario, y un corpus lexicográfico (la BDLex), que se usa en esos mismos trabajos como información complementaria.

b) La redacción de la obra lexicográfica: en la actualidad el *Diccionari descriptiu de la llengua catalana* (DDLC) se encuentra en fase de redacción y muy próximamente se va a publicar a través de Internet la parte de este diccionario ya redactada.

El proyecto DCC se realiza íntegramente en el IEC bajo la dirección de Joaquim Rafel i Fontanals, y ha sido financiado con fondos específicos por el Ministerio de Educación y Cultura y por la CIRIT de la Generalitat de Catalunya.

3. Descripción del CTILC: estructura y composición

El CTILC tiene una extensión global de 52,3 millones. Este volumen de datos textuales da lugar, una vez realizadas las operaciones correspondientes de identificación de unidades léxicas (filtraje de nombres propios y de secuencias no analizables), a una base de datos léxica de 51,2 millones de palabras, que se toma como referencia cuantitativa del corpus en sí. El total de palabras texto que contiene el corpus sen encuentran repartidas, aproximadamente, en un 56% de lengua no literaria y un 44% de lengua literaria (Cf. Fig. 1).

Puede sorprender, tal vez, desde una perspectiva actual, la poca diferencia relativa entre los dos tipos de lengua o, dicho de otro modo, la elevada representación que se da al tipo de lengua literario (un 44%) respecto al no literario (un 56%). De todos modos, cabe tener en consideración que, desde el punto de vista de la "tradición" en la constitución de corpus de referencia "*avant la lettre*" para explotación lexicográfica, en los años en que fue diseñado el CTILC, la tendencia general era justamente la contraria: es decir, privilegiar la lengua literaria por encima de la no literaria. El CTILC invierte esta tendencia en un sentido que, además, se acentuará en los corpus constituidos con posterioridad. No entraré en un análisis detallado de las causas de esta inversión, baste decir que sobre datos objetivos puede evidenciarse la importancia y la diversidad del discurso no literario (científico-técnico, periodístico, etc.) en cualquier repertorio de referencia descriptiva.

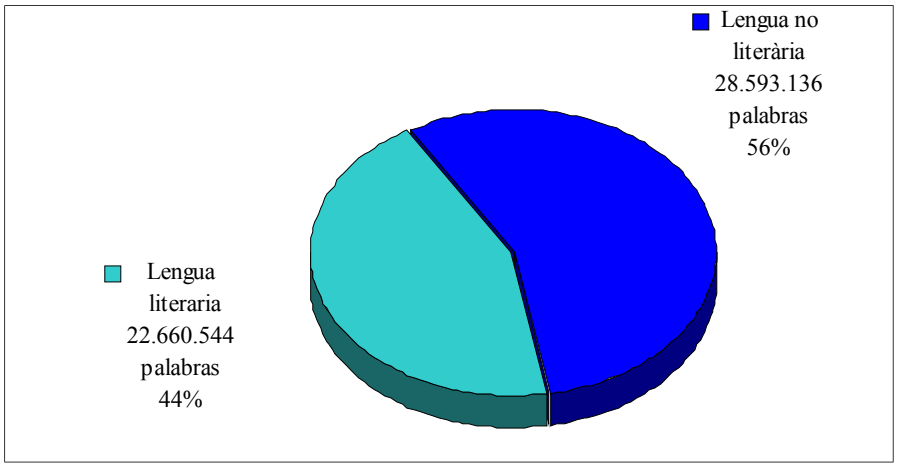


Fig. 1. CTILC: distribución entre lengua literaria y lengua no literaria

El establecimiento de esta división previa del CTILC entre estos dos tipos de lengua, implica también que los criterios de clasificación de los textos se adecuan a cada tipo de lengua. Estos criterios son los siguientes:

- la lengua literaria se divide en cuatro géneros (correspondientes a los cuatro géneros tradicionales de *narrativa, poesía, teatro* y *ensayo*)
- la lengua no literaria se ha estructurado en 10 grupos de base temática:

1 Filosofía.
2 Religión y Teología.
3 Ciencias Sociales.
4 Prensa.
5 Ciencias Puras y Naturales.
6 Ciencias Aplicadas.
7 Bellas Artes. Ocio. Deportes. Juegos.
8 Lengua y Literatura.
9 Historia y Geografía. Biografía.
0 Correspondencia.

Cada uno de estos grupos se divide a su vez en 10 subgrupos de base temática que determinan con una precisión mucho más aproximada la naturaleza interna del texto.

En el cuadro correspondiente a la fig. 2 podemos ver la repartición del total de ocurrencias del corpus en cada uno de los diferentes grupos tipológicos en que se han subdividido la lengua literaria y la lengua no literaria.

Grupo tipológico	Ocurrencias	%
1 Filosofía.	1.730.749	6,05
2 Religión y Teología.	2.922.415	10,22
3 Ciencias Sociales.	5.490.046	19,20
4 Prensa.	3.447.517	12,06
5 Ciencias Puras y Naturales.	2.189.052	7,66
6 Ciencias Aplicadas.	4.422.174	15,47
7 Bellas Artes. Ocio. Deportes. Juegos.	2.735.839	9,57
8 Lengua y Literatura.	2.182.574	7,63
9 Historia y Geografía. Biografía.	3.343.008	11,69
0 Correspondencia.	129.762	0,45
Total NO LITERARIO	**28.593.136**	**55,79**
A Ensayo.	2.996.755	13,22
P Poesía.	2.471.281	10,91
N Narrativa.	13.579.181	59,92
T Teatro.	3.613.327	15,95
Total LITERARIO	**22.660.544**	**44,21**
TOTAL:	**51.253.680**	

Fig. 2. Distribución tipológica

La representación de estos mismos datos puede verse en forma de gráfico en la Fig. 3 y 4 para la lengua no literaria y la lengua literaria respectivamente.

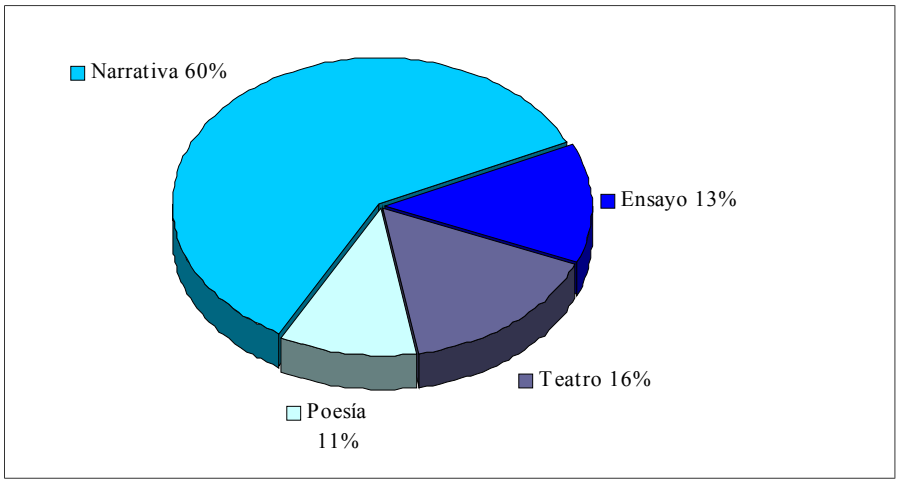


Fig. 3. Lengua literaria: repartición de ocurrencias.

Desde el punto de vista de su cobertura cronológica el CTILC abarca un período de unos 150 años, que se estructura en tres secciones tal como se indica en el cuadro de la Fig. 5. Sin tratarse, pues, de un corpus diacrónico, sí que ofrece una cierta visión evolutiva del catalán contemporáneo en su totalidad y no exclusivamente del catalán estrictamente actual. Esta opción descriptiva, aunque no sea demasiado común en el diseño de los corpus de referencia actuales (que se refieren habitualmente a períodos cronológicos mucho menores) obedece al motivo que el desarrollo de las funciones académicas del IEC y la dimensión social del CTILC hacían necesaria la toma en consideración de un período de una amplia cobertura cronológica, sin que ello supusiera la necesidad de diseñar un corpus que cubriera todos los períodos históricos de la lengua.

Grupo	Subgrupos	Palabras
SECCIÓN I (1833-1873)	4 grupos cronológicos de 10 años	2.260.083
SECCIÓN II (1874-1913)	4 grupos cronológicos de 10 años	8.457.482
SECCIÓN III (1914-1988)	15 grupos cronológicos de 5 años	41.654.379

Fig. 5

Estas divisiones corresponden a tres diferentes períodos de desarrollo del uso social de la lengua bien diferenciados desde el punto de vista histórico.

Las figs. 6 y 7 nos muestran, respectivamente, cual es la repartición cronológico-tipológica del CTILC. En ellas se puede ver claramente qué grado de representación se ha otorgado en el corpus a cada una de sus divisiones cronológicas. Ha primado, como puede verse, la representación de los períodos cronológicos de la lengua más cercanos a nuestros días, por tanto se trataba de representar en grado máximo los usos lingüísticos que se mantienen vigentes en la actualidad.

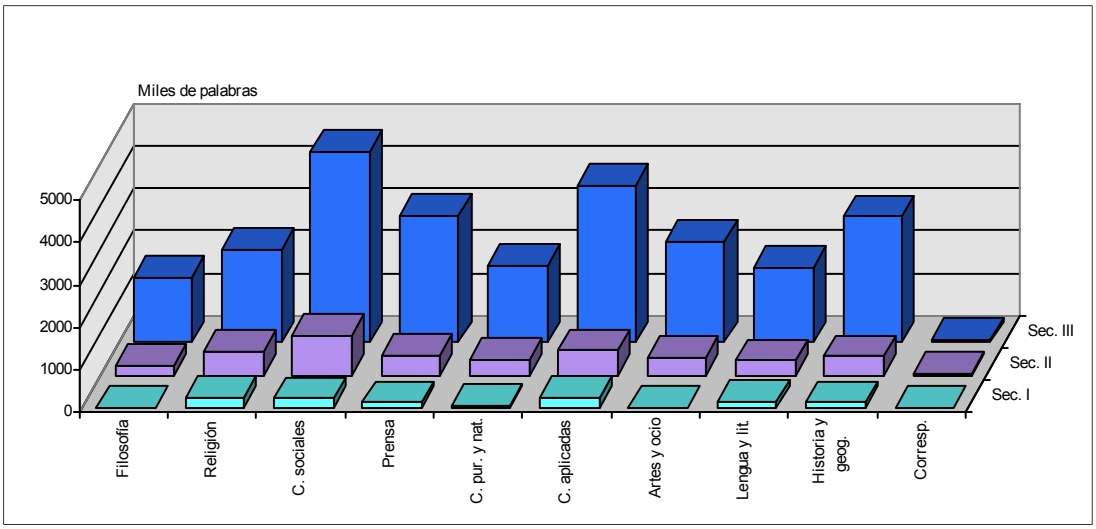
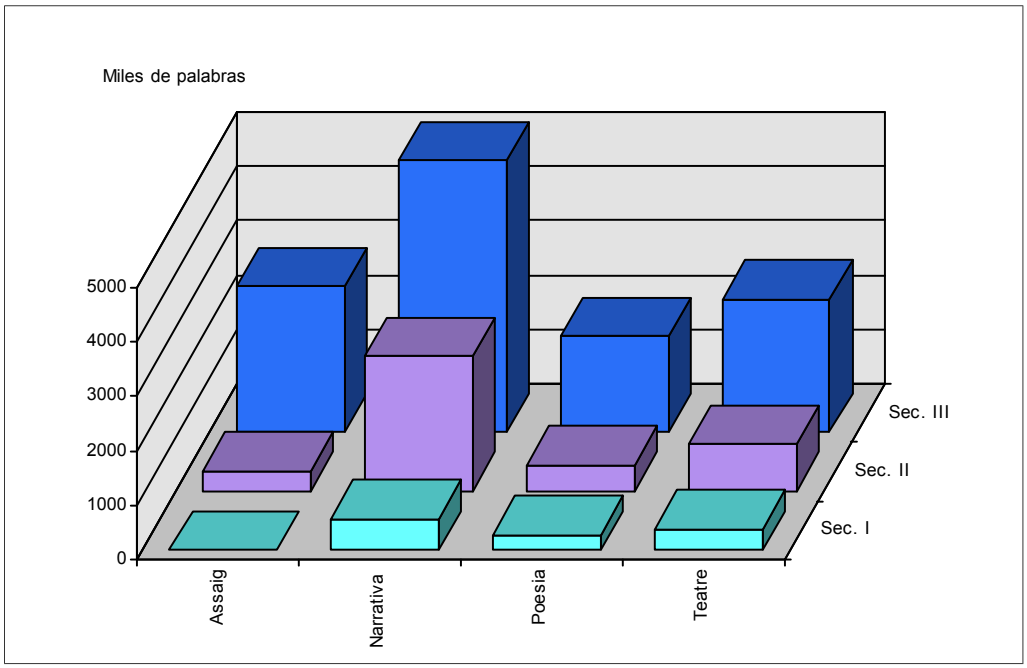


Fig. 6. Lengua no literaria: repartición cronológico-tipológica.

Fig 7. Lengua no literaria: repartición cronológico-tipológica.



El número de obras que forman parte del corpus es de 3.299 (1.011 correspondientes a la lengua literaria y 2.288 correspondientes a la lengua no literaria).

4. Constitución del CTILC: representación textual

Las obras han sido incorporadas al CTILC en su versión íntegra. En muy pocos casos se han introducido parcialmente, por su extensión fuera de lo común, algunas obras pertenecientes al subcorpus no literario, de las que se seleccionaron únicamente

determinados capítulos. La edición tomada como edición de referencia es, salvo en poquísimas excepciones debidamente justificadas, la primera, y la publicación en papel se ha usado para la referenciación de la información del corpus en términos de página, línea, y número de orden de la palabra dentro de la línea.

En un porcentaje muy elevado, los textos del CTILC fueron introducidos manualmente, a la vez que los códigos de representación textual necesarios para determinados procesos de constitución de la base de datos textual y para la reconstrucción del texto a partir de ésta. El tipo de obras seleccionadas, y el estado de desarrollo tecnológico (muy inferior al actual) de los sistemas de lectura óptica determinó que el uso de estos fuera limitado únicamente a los textos cuya extensión y características físicas permitían optimizar las tareas de entrenamiento de los reconocedores de caracteres. Por otra parte, la práctica inexistencia (en el período cronológico cubierto por el CTILC) de textos en formato electrónico procedentes de editoriales, periódicos, etc., no hizo posible la reutilización de material textual. Con todo, sí que se incorporaron al CTILC, con las adaptaciones de codificación correspondientes, textos en soporte magnético introducidos en experiencias previas de constitución de corpus limitados, como las del proyecto *Prospecció automatitzada de textos* catalans, de la Universidad de Barcelona.

El sistema de representación textual del CTILC, aplicado en el momento del traslado a soporte electrónico de los textos, integra tres tipos diferenciados de elementos de marcaje:

a) Transcripción desambiguada de carácteres textuales: espacios en blanco que forman parte integrante de palabras (*a priori*); diferentes tipos de guiones (*estat-generalitat, penja-robes, agafar-me*). Un caso especial de este tipo de elementos son los códigos que nos permiten añadir información necesaria para la interpretación textual (*mal p…*) o eliminar información no pertinente para el procesamiento léxico (*ex-tra-or-di-na-ri*).

b) Marcaje de informaciones tipográficas: cursiva, negrilla, versalitas.

c) Marcaje de informaciones lógicas para la realización de acciones específicas sobre cada uno de los tipos de información del corpus. Estas informaciones se pueden organizar entorno a dos criterios de clasificación: por una parte, su ámbito de aplicación (que puede ser o bien una sola palabra o un fragmento textual de un párrafo como máximo); por otra parte, el hecho de que determinen que la información marcada sea o no pertinente para la elaboración del diccionario. En este último caso, los códigos que determinan la exclusión de información de cara a la explotación léxica del corpus son el de *nombre propio* y el de *información no analizable*. En la fig. 8 puede verse la clasificación de estos elementos.

Información	Ámbito	Excl.
Acotación teatral	Texto	No
Fragmento en verso	Texto	No
No analizable	Texto	Sí
Nombre propio	Palabra	Sí
Palabra de otra lengua	Palabra	No
Palabra en mayúscula	Palabra	No
Sigla o abreviatura	Palabra	No
Título	Texto	No

Fig. 8. Codificación textual del CTILC

5. Constitución del CTILC: lematización

El CTILC es un corpus completamente lematizado (anotado y desambiguado en lo referente a la información morfosintáctica). Por sus características especiales, el proceso de lematización semiautomática ha representado también la validación manual de todos los datos lematizados. Por esta razón puede decirse que el CTILC es un corpus altamente fiable para la extracción de datos lingüísticos y que es adecuado no solamente para la actividad lexicográfica sino también para la investigación lingüística en general.

La relación lema/forma con que se etiqueta la información léxica del corpus implica, desde el punto de vista de la organización de los datos lingüísticos, dos operaciones complementarias:

a) Agrupación de todas las posibilidades de aparición de un mismo elemento léxico en el discurso.

b) Desambiguación de las formas homógrafas.

Con frecuencia, en proyectos de constitución de corpus, estas dos operaciones aparecen disociadas y suelen corresponder a dos ciclos diferentes de anotación de los datos. En el caso del CTILC el esquema del proceso de lematización los realiza conjuntamente.

El proceso de lematización, además de relacionar las diferentes formas gráficas de una misma serie flexional, agrupa también las variantes gráficas que pueden aparecer en los textos. Por otro lado, se codifican también como formas los derivados apreciativos (diminutivos, aumentativos, peyorativos, intensivos, marcados con un código morfológico comenzado por "D"), para los cuales no se crean nuevos lemas. Hay que tener el cuenta también que cuando una palabra aparece usada metalingüísticamente en un texto se codifica como tal y se clasifica como una forma diferenciada del resto.

La lista de etiquetas categoriales de los lemas es la siguiente:

A	Adjetivo
AF	Adjetivo femenino (*clàudia, versaleta*)
AI	Adjetivo invariable de género (*benvolent*)
AII	Adjetivo invariable de número (*beix*)
AIP	Adjetivo invariable plural (*sengles, senars*)
AM	Adjetivo masculino (*núvol, senglar*)
AMP	Adjetivo masculino plural (*alisis*)
AN	Adjetivo numeral (*cinc*)
AP	Adjetivo plural (*ambdós*)
AR	Artículo
AV	Adverbio
C	Conjunción
CT	Contracción (*del*)
F	Nombre femenino
FP	Nombre femenino plural (*vacances, ulleres*)
FS	Nombre femenino singular (*viu-viu*)
I	Interjección (*ui, clin clon, vejam*)
IFX	Infijo
L	Palabra ligada (*antuvi, palpentes*)
LA	Locución alóctona (*a priori*)
M	Nombre masculino
MF	Nombre masculino o femenino (*mar, color*)
MP	Nombre masculino plural (*afores*)
MS	Nombre masculino singular (*endemà, fra*)
NC	No codificado (*rai, stella maris, cf, £d glucosa*)
P	Pronombre

PFX	Prefijo
PO	Preposición
SFX	Sufijo
SIG	Sigla, acrónimo o abreviatura de más de una palabra
V	Verbo transitivo e intransitivo
VA	Verbo auxiliar
VI	Verbo intransitivo
VIA	Verbo intransitivo y auxiliar
VIP	Verbo intransitivo y pronominal
VP	Verbo pronominal
VT	Verbo transitivo
VTP	Verbo transitivo y pronominal
VVP	Verbo transitivo, intransitivo y pronominal

Cada ocurrencia se marca también con una etiqueta morfológica que se refiere, fundamentalmente, a sus características flexivas y en ocasiones a sus modalidades apreciativas (etiquetas D), o a su modo de aparición en el discurso como palabra citada (MET). La lista de etiquetas morfológicas para las formas es la siguiente:

a) Flexión nominal

MS	Masculino singular
FS	Femenino singular
MP	Masculino plural
FP	Femenino plural
DMS	Apreciativo masculino singular
DFS	Apreciativo femenino singular
DMP	Apreciativo masculino plural
DFP	Apreciativo femenino plural
S	Singular
P	Plural
DS	Apreciativo singular
DP	Apreciativo plural
D	Apreciativo
MET	Úso metalingüístico

b) Flexión verbal

IF	Infinitivo
GE	Gerundio
RMS	Participio masculino singular
RFS	Participio femenino singular
RMP	Participio masculino plural
RFP	Participio femenino plural
XPI	X persona presente de indicativo
XII	X persona imperfecto de indicativo
XPT	X persona perfecto
XFU	X persona futuro
XPS	X persona presente de subjuntivo
XIS	X persona imperfecto de subjuntivo
XCO	X persona condicional
XIM	X persona imperativo

Las figs. 9 y 10 nos muestran de forma esquemática la naturaleza de la relación lema/forma.

Forma	Freq	Lema	CGram	CMorf	Freq
força	17.856	**força**	AII	—	1110
				MET	2
		força	AV	—	3290
				MET	1
		força	F	S	13339
				MET	11
		forçar	VT	3PI	102
				2IM	1

Fig. 9. La relación lema/forma en la forma gráfica *força*.

En la figura 9 pueden verse las diferentes posibilidades de lematización de la forma gráfica *força*, que aparece 17.856 veces a lo largo del corpus. Las dos columnas centrales nos indican el lema a que puede pertenecer esta forma gráfica, mientras que las dos columnas de la derecha indican los diferentes valores morfológicos (y su frecuencia) con que la forma *força* aparece ligada a los distintos lemas.

Lema	CGram	Forma	CMorf	Freq
força	F	força	S	13.339
		fòrça	S	46
		forsa	S	2.917
		fórsa	S	25
		fòrsa	S	115
		forse	S	8
		fòrse	S	1
		forssa	S	9
		forza	S	3
		forças	P	48
		forces	P	5.556
		fòrces	P	8
		forçes	P	71
		fòrçes	P	3
		forsas	P	739
		fórsas	P	2
		forses	P	198
		fórses	P	2
		fòrses	P	6
		forssas	P	5
		forsses	P	2
		forzas	P	1
		forçarra	DS	2
		força	MET	11
		forces	MET	1
		forsa	MET	1

Fig. 10. Las formas del lema **força** *f.*

La figura 10, por otra parte, nos muestra las diferentes formas que el sustantivo femenino *força* tiene asociadas como resultado del proceso de lematización del corpus, con la frecuencia de cada una en la columna de la derecha.

Existen dos tipos de lemas que se relacionan jerárquicamente: los lemas principales y los secundarios. Un lema principal puede relacionarse con *n* lemas secundarios. Esta distinción se aplica en aquellos casos en que la variación va más allá de la puramente gráfica (sin que tenga nada que ver con ninguno de los casos precedentes), y comporta una diferencia en la estructura fonológica o morfológica de la palabra, que puede parecer suficiente para determinar lemas diferentes, pero con una similitud suficientemente grande como para mantener algún tipo de relación lógica que permita tratarlos, en caso conveniente, de un modo agrupado. Las figuras 11 y 12 nos muestran dos ejemplos de relación entre lemas principales y secundarios.

LEMA	RANGO	FORMA
DONCS C	Principal	donç
		donchs
		dónchs
		dònchs
		doncs
		dóncs
		dons
		donchs MET
		doncs MET
		dòngks MET
		dòngs MET
		dons MET
		dòns MET
DONC C	Secundario	donc
		donch
DONCES C	Secundario	donças
		donsas
DONQUES C	Secundario	doncas
		donques
DÒS C	Secundario	dos
		dòs
		dôs
		dòs MET

Fig. 11.Tipos de lemas

LEMA	RANGO	FORMA
ESMORZAR M	Principal	esmorçar S
		esmorsà S
		esmorsar S
		esmorzar S
		esmorçars P
		esmorsars P
		esmorzars P
		esmorzar MET
ALMORZAR M	Secundario	almorzar S
ARMOSAR M	Secundario	armosar S
		armosarot DS

Fig. 12. Tipos de lemas

Como resultado del proceso de lematización, los datos frecuenciales del CTILC pueden darse como se muestra en la figura 13, que pone en relación la extensión absoluta de cada grupo tipológico del corpus con el número de lemas que aparecen en cada grupo. De este modo puede observarse a grandes rasgos el grado de "riqueza léxica" (renuncio a definir esta expresión desde el punto de vista léxico) de cada una de las

divisiones del corpus.

Grupo tipológico	Ocurrencias	Lemas
1 Filosofía.	1.730.749	24.932
2 Religión y Teología.	2.922.415	30.204
3 Ciencias Sociales.	5.490.046	44.983
4 Prensa.	3.447.517	44.211
5 Ciencias Puras y Naturales.	2.189.052	39.997
6 Ciencias Aplicadas.	4.422.174	50.777
7 Bellas Artes. Ocio. Deportes. Juegos.	2.735.839	36.312
8 Lengua y Literatura.	2.182.574	35.183
9 Historia y Geografía. Biografía.	3.343.008	37.602
0 Correspondencia.	129.762	9.044
Total NO LITERARIO	**28.593.136**	**118.700**
A Ensayo.	2.996.755	37.342
P Poesía.	2.471.281	37.481
N Narrativa.	13.579.181	68.083
T Teatro.	3.613.327	34.423
Total LITERARIO	**22.660.544**	**86.049**
TOTAL:	**51.253.680**	**148.627**

Fig. 13

Vale la pena, aunque sea superficialmente, poner de relieve el valor que datos como los que se muestran en la figura 13 pueden tener en el diseño de corpus más "equilibrados" desde el punto de vista de la variedad léxica que pueden aportar sus diferentes particiones tipológicas. A partir de la relación entre el número de lemas de cada división tipológica del CTILC y del número total de ocurrencias que representa dicha partición, podemos sacar conclusiones de gran interés para el diseño futuro de corpus de referencia e, incluso, para la actualización permanente del CTILC.

5. Usos y accesibilidad del CTILC

Paralelamente a la constitución del CTILC, se diseñó la base de datos y el interfaz de acceso al corpus, que se usan actualmente en los trabajos de redacción del diccionario descriptivo que se desarrolla en la segunda fase del proyecto DCC. Los requisitos que tenía que reunir esta base de datos fueron satisfechos en su totalidad: independencia del volumen de datos, convertibilidad de la información del corpus a formalismos estándar, implementación de los módulos de extracción de información colocacional y de recuperación de la información contextualizada, etc.

Posteriormente al diseño de estas herramientas de uso interno, se ha dado acceso al CTILC a través de Internet a partir de un interfaz de acceso diseñado específicamente como componente del *Portal de dades lingüístiques* (PDL) del IEC (http://www.iecat.net, vínculo PDL: http://pdl.iecat.net).

Uno de los resultados destacables a que ha dado lugar CTILC es la publicación (en libro CD-ROM) del *Diccionari de freqüències* (3 volúmenes publicados entre 1996 y 1998). Este diccionario recoge no sólo las informaciones frecuenciales básicas del corpus, sino también informaciones referentes al grado de repartición de los lemas de acuerdo con los diferentes grupos tipológicos de los subcorpus literario y no literario.

Los datos estadísticos que contiene el diccionario de frecuencias constituyen una aportación de gran valor para el estudio de la estructura cuantitativa del léxico catalán.

La combinación de la frecuencia propiamente dicha con los valores de *distribución* (es decir, el parámetro que nos indica si la repartición de los lemas a lo largo de las distintas divisiones tipológicas del corpus es más o menos uniforme) de los lemas permite extraer conclusiones de gran valor sobre la importancia relativa de cada lema en relación con la totalidad del vocabulario. Una de las aplicaciones internas más destacables de los datos de este diccionario ha sido la selección de la nomenclatura para el diccionario descriptivo.

El CTILC se ha utilizado y se utiliza en la actualidad como fuente de información empírica para un buen número de trabajos lingüísticos de naturaleza muy diversa que necesitan el concurso de datos lingüísticos. Para asegurar en un grado máximo la multifuncionalidad y las posibilidades de reutilización del CTILC, se han diseñado los módulos de conversión de la información del CTILC a los formatos de representación estándar establecidos por EAGLES y adaptados en el marco del proyecto PAROLE, en el que nuestro grupo participó como responsable de la ejecución del corpus para el catalán.

6. Conclusiones y reflexiones

En esta presentación general del CTILC he pretendido hasta ahora definir con cierto grado de detalle los aspectos de diseño y estructura. Afortunadamente, el estado actual de desarrollo y de conocimiento en relación al diseño y a la construcción de corpus es muy diferente al existente en el momento en que el CTILC fue diseñado y, en buena parte, constituido.

En la actualidad la constitución de corpus es mucho más factible, los costes de financiación de este tipo de proyectos se han reducido considerablemente, así como también los períodos de ejecución y el esfuerzo que requiere la constitución de corpus de dimensiones medianas o grandes. Esta coyuntura objetivamente más favorable se debe a factores de índole muy diversa, entre los que podemos destacar:

a) El desarrollo de herramientas de constitución (etiquetadores automáticos) y de explotación (*browsers*) que permiten concentrar esfuerzos en el proceso de obtención de los textos y no tanto en su tratamiento posterior.

b) La cantidad y la diversidad de los textos existentes en soporte electrónico, que permiten la reunión de grandes volúmenes de datos sin necesidad de grandes esfuerzos dedicados al traslado a soporte electrónico.

c) Los estándares de codificación lingüística y de representación desarrollados en los últimos años, que permiten optimizar las posibilidades de intercambio y de adaptación de textos, y a su vez facilitan el trabajo conceptual de diseño de los procesos de constitución de corpus y la utilización de herramientas genéricas.

d) La necesidad reconocida por la lingüística descriptiva de contar con repertorios representativos de datos lingüísticos que permitan la superación de las observaciones introspectivas como único elemento decisivo en el aporte de datos.

e) La necesidad de las lenguas (especialmente de las minoritarias) de incorporarse a la denominada sociedad de la información, lo que implica mayores necesidades de recursos lingüísticos que hagan posible los desarrollos en el campo de las industrias de la lengua que puedan asegurar esta presencia.

Finalmente, y una vez constatadas estas ventajas objetivas en el panorama de los proyectos de constitución de corpus, cabe preguntarse hasta qué punto la reunión de

estos factores coyunturales incide, a su vez, en las características de diseño de los corpus y si modifica de algún modo algunas de sus características principales. Algunas de estas características que habitualmente se asocian a los corpus de referencia pueden ser la multifuncionalidad, la reutilización y convertibilidad, la representatividad y la fiabilidad. Son totalmente independientes estas características de los procesos de constitución de corpus? O pueden verse adaptadas, redefinidas, e incluso alteradas por los cambios que se han producido en las posibilidades técnicas y tecnológicas?

En mi opinión, la incidencia de la posibilidades actuales es distinta en relación a cada uno de estos aspectos. Los aspectos de reutilización y convertibilidad de la información, por ejemplo, se han visto enormemente favorecidos por la aparición de estándares de codificación que han sido aplicados mayoritariamente. En el extremo contrario, quizás, podríamos señalar que la utilización de etiquetadores automáticos representa, al menos en la actualidad, un descenso del grado de fiabilidad lingüística (o, en todo caso, la necesidad de plantear una redefinición del concepto de fiabilidad de un corpus).

La incidencia de las facilidades que se pueden encontrar hoy en día en lo referente a reutilización de textos electrónicos para la creación de corpus puede incidir también notablemente en la noción de representatividad de los corpus de referencia. Cabe plantearse hasta qué punto el tipo de textos más accesible en soporte electrónico (habitualmente prensa periódica, documentación de páginas web, etc.) puede determinar o limitar la variabilidad (y por lo tanto indirectamente la representatividad) de los corpus de referencia que se construyan en un futuro inmediato. En cierto modo, y simplificando mucho, podría decirse que la facilidad en la adquisición de material textual podría llevar a replantear (de un modo que a mi entender resultaría contraproducente) a adaptar a esa nueva realidad el concepto de representatividad de los corpus.

Con esta observación no quiero, naturalmente, cuestionar la necesidad del salto cualitativo que se ha producido en los últimos años. Es obvio que las perspectivas actuales son favorables para la constitución de corpus de referencia e, incluso, de corpus *monitor* (corpus de actualización permanente que pretenden dar cuenta de los fenómenos de variación y cambio asociados al uso de la lengua) para dar cuenta de los procesos de cambio lingüístico que se producen en nuestras sociedades. Eso de por sí ya es un factor enormemente positivo. Es responsabilidad, pues, de las instituciones encargadas de llevar a cabo los proyectos de constitución de recursos lingüísticos el dar forma a los requerimientos de cada corpus de referencia, sin dejarse llevar por el camino fácil de una falsa eficiencia que pueda conducirles a resultados desviados desde el punto de vista de los datos lingüísticos recogidos, pero aprovechando todas las posibilidades actuales.