

"Prospecció Automatitzada de Textos Catalans"

El proyecto de investigación "Prospecció Automatitzada de Textos Catalans" (PATC), cuyas fases iniciales se están llevando a cabo en la Universidad de Barcelona, tiene sus orígenes en el deseo de crear un centro dedicado a la investigación lingüística asistida por ordenador. El Anteproyecto para la creación de un centro de lingüística automática (abril 1975) y la Memoria para la creación de un centro de lingüística automática en la Universidad de Barcelona (enero 1976), elaborados conjuntamente por el profesor Santiago Mollfulleda y por el que suscribe, son los primeros documentos redactados en relación con esta idea, que desde tiempo atrás estaba en la mente de sus autores.

Con motivo del primer intento de poner en marcha este proyecto, tuvimos que concretar mucho y delimitar nuestros objetivos. En relación con ello debemos citar por una parte la celebración en Barcelona el 26 y 27 de Mayo de 1977 de una mesa redonda sobre informática y lingüística a la que presentaron comunicaciones, aparte de los profesores E. Moreu, Juan Vernet, M^a Teresa Cabré y M^a Angels Vidal de Barcelona, prestigiosos investigadores de diversos centros europeos (L. Delatte, de Lieja, Mario Alinei, de Utrecht, M. Tournier, de St. Cloud, P. Imbs, de Nancy, G. Colon, de Basilea) --véase el nº 1 de la serie Informática y lingüística (Mesa redonda sobre informática y lingüística. Barcelona, 26 y 27 de Mayo de 1977), publicación ciclostilada editada por la "Fundación para el desarrollo de la función social de las comunicaciones" (FUNDESCO); y, por otra parte, debemos aludir a la memoria Prospecció Automatitzada de Textos Catalans (504 págs.) redactada según el esquema para memorias de trabajos científicos y proyectos de investigación de FUNDESCO cuando intentábamos llevar adelante el proyecto con el patrocinio de esta fundación y con el concurso de diversas instituciones catalanas (el equipo redactor de la memoria estuvo constituido por los profesores M^a Teresa Cabré, Carmen González, Teresa Gracia, Santiago Mollfulleda, Mercè Otero, Pere Quetgles y Joaquín Rafel, y por el informático Valentín Ibáñez). Esta po-

sibilidad no llegó a cuajar por diversos motivos, pero en aquel momento habíamos conseguido pasar de un proyecto inicial vago e impreciso a otro de un notable grado de concreción, descrito de un modo tan detallado que permitía el inicio inmediato de su ejecución.

Los medios de financiación no llegaron hasta el año 1980 en que se convocaron por primera vez las Ayudas a la investigación por la Universidad de Barcelona, en principio para proyectos de un año de duración. En los años sucesivos se han ido concediendo ayudas para diversas fases del proyecto.

Sin perder de vista el objetivo general que nos habíamos trazado al principio, el proyecto específico que ha resultado tiene por finalidad fundamental la de crear un archivo de textos —en principio literarios— de la lengua catalana sobre soporte magnético automatizable con la condición básica de mantener toda la información que se contiene en el documento llamado Información fuente (IF) (edición moderna, incunable, manuscrito, etc.), para su ulterior utilización en la investigación lingüística. Además, como objetivos accesorios, consideramos la creación de programas generales para la explotación científica de los textos almacenados. La gran mayoría de los textos están escritos en lengua catalana, pero algunos de ellos lo están en latín; se trata de textos medievales que tienen un gran interés para la historia de la lengua catalana.

Por lo que respecta a los textos en latín, se han introducido en soporte magnético y se han tratado informáticamente los Usatges de Barcelona y el Cançoner eròtic de Ripoll; para estos textos se han utilizado los servicios del Laboratoire pour l'analyse statistique des langues anciennes de la Universidad de Lieja.

En cuanto a los textos propiamente en lengua catalana, en estos momentos tenemos completamente introducidos y corregidos los textos siguientes:

Tirant lo Blanc (extensión: 3.000.000 de caracteres)
(IF: facsímil del incunable de 1492,
propiedad de la Hispanic Society of
America)

El Quadern Gris de Josep Pla (ext.: 1.600.000 car.)

Se está terminando de introducir la Obra narrativa completa de Joaquim Ruyra (ext.: 2.000.000 de caracteres) y se ha comenzado la introducción de las Memòries de Josep M^e de Sagarra (ext.: 2.500.000 caracteres). Para el año 1984 se tiene programada la introducción de la obra narrativa completa de Víctor Català (ext. 3.000.000 de car.) y de Narcís Oller (4.000.000 de car.).

La introducción de la información sobre soporte magnético automatizable se hace a través de una estación de datos IBM 3741, en disco flexible. El almacenamiento definitivo se hace en cinta magnética.

Aparte de la introducción en sí misma, una de las labores fundamentales del equipo de trabajo ha sido la elaboración de los distintos documentos relacionados con el proceso informático, empezando por las codificaciones e instrucciones de entrada de datos y de corrección, y terminando por los programas para la obtención de distintos índices, aparte de otros trabajos que se están realizando.

Los índices que pensamos obtener de todo texto introducido —ya existen los de El Quadern Gris— son:

1º Índice de palabras ordenadas alfabéticamente con indicación de las frecuencias absoluta y relativa de cada una.

2º Índice de palabras ordenadas por orden decreciente de frecuencias.

3º Índice de palabras ordenadas alfabéticamente, acompañadas de la indicación de las referencias de su ubicación en el texto (página, línea, número de orden de la palabra en la línea —eventualmente, columna, capítulo, etc.).

4º Índice inverso (palabras ordenadas por la parte final), con indicación del número de finales distintos de uno, dos y tres caracteres).

Frecuentemente, en vez —o además— del nº 3, se obtendrá lo que se conoce con el nombre de concordancias, es decir, además de la información que contiene el índice nº 3, y junto a la referencia de cada ocurrencia, un fragmento del texto para cada

una de las apariciones de cada palabra.

Para todos estos índices hemos elaborado los programas adecuados y los resultados son listados sobre papel y almacenados en cinta magnética (una muestra de ellos, por lo que respecta a la obra El Cuadern Gris, son adjuntados a este informe). Existe el proyecto de hacer una edición de ellos en microficha para utilizar a través de una lectora-copiadora.

Al margen de estos resultados particulares, o de otros cualesquiera, el texto queda almacenado, tanto en su primitiva forma, que reproduce la de la Información fuente, como en forma de lo que hemos llamado base de datos. Esta base de datos consta de un registro de 26 campos para cada palabra de la obra; en él se encuentra, aparte de la misma palabra, toda la información necesaria para su adecuada referenciación, catalogación, ordenaciones diversas, etc. A la base de datos se puede acudir para obtener cualquier información que no esté contenida en los índices editados.

Además del desarrollo de los aspectos mencionados del proyecto, constituyen también parte del mismo los trabajos que se están llevando a cabo para realizar una lematización automatizada de los textos. La asociación de todas las formas flexivas a una única forma representativa de la palabra (LEMA), y la separación de formas homógrafas, asociándolas a lemas distintos, es uno de los caballos de batalla del tratamiento exhaustivo de textos. Nuestro proyecto de lematización comprende un programa —todavía no completamente elaborado— que obtendrá un lema (o varios posibles lemas) después de someter cada palabra del texto a una serie de substituciones de su parte final, alternadas con la confrontación del resultado de cada operación con un diccionario almacenado sobre soporte magnético automatizable; los resultados de todas estas operaciones serán sometidos a un investigador para que los dé por buenos o para que elija entre varios lemas propuestos. En el momento actual tenemos sobre soporte magnético un diccionario de 50.000 entradas con sus indicadores gramaticales normalizados y elaborados los algoritmos de análisis morfológico para el tratamiento de plurales de sustantivos,

femeninos y plurales de adjetivos, flexión verbal, diminutivos y superlativos sintéticos.

Por lo que se refiere a los aspectos materiales (hardware), —aparte de la utilización del centro de Lieja para los textos en latín, a que ya nos hemos referido— unas fases del proyecto se han desarrollado en un centro de cálculo comercial, cuando las condiciones del Laboratorio de Cálculo de la Universidad no reunían los requisitos mínimos exigidos para un trabajo de este tipo; desde hace un año, tras la instalación de una máquina IBM 4341 y una reorganización del Laboratorio de Cálculo, con instalación de terminales en las Facultades, etc., hemos empezado a utilizar estos servicios. El principal problema que tenemos pendiente es el de la obtención de una impresora adecuada para los resultados en papel; de momento, tal como se ve en los especímenes que reproducimos junto a este documento, hemos debido recurrir a una ficción para representar los conceptos más imprescindibles, con impresión en interlínea (acento grave, acento agudo, diéresis, mayúscula), o con sobreimpresión (cedilla). En los listados de trabajo (correcciones, etc.) estos y otros conceptos vienen representados por su referencia interna y se trabaja con listas de correspondencias, pero en los resultados para ofrecer a personas externas al propio proyecto, nos pareció que, aunque fuera provisionalmente, debíamos elaborar los programas que permitieran una lectura no mediatizada por instrucciones demasiado complejas. De todas formas, no olvidamos el problema, y estamos colaborando con la dirección del Laboratorio de Cálculo en el sentido de estimular la instalación de una impresora adecuada que permita representar cualquier tipo de signo especial.

Ni que decir tiene que el futuro de un programa de esta naturaleza depende en gran medida de las posibilidades de financiación, y que el modo como hasta ahora se ha venido desarrollando este proyecto no asegura su continuidad. Esperamos —y trabajamos en ello— poder consolidar este aspecto fundamental y poder continuar ofreciendo información sobre futuros logros.

Joaquim RAFEL i FONTANALS

1. Índice de palabras ordenadas alfabéticamente con indicación de las frecuencias absoluta y relativa de cada una.

JARDEN UNIS		MUT		274	
MOT	F. ABS	F. REL	MOT	F. ABS	F. REL
POSAVA	24	0,009	POSSEIR	8	0,003
POSAVEM	3	0,001	POSSEIT	1	0,000
POSAVEN	5	0,001	POSSESSIC	9	0,003
POSEM	6	0,002	POSSESSIVES	1	0,000
POSEN	13	0,005	POSSIBILISME	1	0,000
POSÉS	4	0,001	POSSIBILITAT	23	0,009
POSSESSIN	1	0,000	POSSIBILITATS	15	0,005
POSEU	3	0,001	POSSIBLE	101	0,039
POSI	5	0,001	POSSIBLEMENT	1	0,000
POSICIÓ	43	0,016	POSSIBLES	8	0,003
POSICIONS	6	0,002	POST	2	0,000
PCSIN	1	0,000	POSTA	9	0,003
POSIS	1	0,000	POSTAL	3	0,001
PCSITIU	9	0,003	POSTAL	1	0,000
POSITIUS	2	0,000	POSTALS	1	0,000
POSITIVA	16	0,006	POSTERIOR	7	0,002
POSITIVAMENT	11	0,004	POSTERIORITAT	2	0,000
POSITIVES	6	0,002	POSTERIORES	4	0,001
POSITIVISME	1	0,000	POSTES	2	0,000
PCSG	13	0,005	POSTISSA	1	0,000
POSSEIX	9	0,003	POSTRES	7	0,002
POSSEIXEN	1	0,000	POSTULEN	1	0,000
POSSEIXI	1	0,000	POSTURES	1	0,000
POSSEI	1	0,000	POSTURETES	1	0,000
POSSEIA	1	0,000	POT	225	0,088
POSSEIDES	1	0,000	POTA	1	0,000
POSSEIDOR	1	0,000	POTABLE	1	0,000
POSSEIDORES	1	0,000	POTABLES	2	0,000
POSSEIEN	2	0,000	PCTACURT	1	0,000

2. Índice de palabras ordenadas por orden decreciente de frecuencias.

39		
MOT	F.ABS	F.REL
ESPERANÇA	11	0,004
ESSERS	11	0,004
ESTAVEN	11	0,004
ESTRIS	11	0,004
ESTUDIS	11	0,004
EXACTA	10	0,003
EXAMEN	10	0,003
EXAMINAR	10	0,003
EXCEPCIONAL	10	0,003
EXERCICIS	10	0,003
EXPLICACIÓ	10	0,003
EXPOSICIÓ	10	0,003
EXQUISIDA	10	0,003
EXQUISIDES	10	0,003
EXQUISIT	10	0,003
FANAL	10	0,003
FATXENDA	10	0,003
FEBRE	10	0,003
RASTRE	10	0,003
FINAL	10	0,002
FÍSICAMENT	10	0,003
ABUNDANCIA	10	0,003
FORASTERS	10	0,003
ACLAPARAT	10	0,003
RELACIONS	10	0,003
ACTIVA	10	0,003

3. Índice de palabras ordenadas alfabéticamente acompañadas de las referencias de su ubicación en el texto (página, línea y número de la palabra dentro de la línea en la Información fuente).

QUADERN GRIS	LOCALITZACIONS														
401															
DESTACAVA	579	4	3,	850	17	10,									
DESTAPEN	299	32	6,												
DESTAQUEN	664	26	5,												
DESTEIXIR	406	17	10,												
DESTEMPRADA	333	5	7,												
DESTENYIDES	503	12	4,												
DESTENYIT	542	27	2,												
DESTI	295	12	7,	692	36	6,									
DESTIL·LACIÓ	404	17	6,												
DESTIL·LAT	511	12	4,												
DESTIL·LO	429	17	5,												
DESTINADA	316	5	4,	326	18	1,	365	22	5,	455	6	1,	554	16	6,
	808	33	6,	843	34	5,	849	12	1,						
DESTINADES	274	29	6,	535	5	6,	854	33	6,						
DESTINAREN	431	15	1,												
DESTINAT	324	11	4,	390	32	4,	474	18	7,	741	30	8,	838	23	1,
DESTINATS	356	3	7,	569	39	6,	814	19	1,						
DESTORBADA	182	30	5,												
DESTRALEJAR	394	37	2,												
DESTRALS	399	18	8,												
DESTREMPADA	550	23	9,												
DESTREMPADES	503	34	3,												
DESTRIAR	761	13	3,												
DESTRUSSA	594	20	6,												
DESTROSSAT	572	27	8,												
DESTRUCCIÓ	220	39	9,	295	21	7,	303	35	2,	456	15	1,	456	16	4,
	483	32	2,	526	28	8,	817	11	3,	819	38	7,			
DESTRUEIX	746	10	2,												
DESTRUÏDA	586	8	4,												

4. Índice inverso (palabras ordenadas alfabéticamente por su parte final, con indicación numérica de los finales distintos de los tres, dos y un caracteres).

CUADERN GRIS

FA---XXX---FA---XX

MOT-INV

FA---XXX---FA---XX

MOT-INV

24

		SORRUDA				GALLOFA
		MOL SUDA				SOFA
-UDA	38	-DA	877		-OFA	4
		PANACEA				DESFA
		OCEA			-SFA	2
-CEA	2					SATISFA
		IDEA				BUFA
-DEA	1					BUFA
		FARMACPEA				
		EUROPEA				BALDUFA
-PEA	2					ESTUFA
		AREA			-UFA	4
		CREA			-FA	24
		MENYSPREA				EMBRIAGA
		VERBORREA				BALIGA-BALAGA
-REA	4					ARGELAGA
		ODISSEA				PLAGA
		NAUSEA				AMAGA
-SEA	2	-EA	11			PASTANAGA
-FA	1					PAGA
		EMBAFA				APAGA
		AGAFA				NAUFRAGA
		AGAFA				NISSAGA
-AFA	4	GIRAFA				VAGA
-EFA	1	REFA				DIVAGA
		RIFA				MANYAGA
		TARIFA			-AGA	13
		TIFA				CEGA
		REVIFA				ENCEGA
-IFA	4					OFEGA
		PELFA				MARFEGA
-LFA	2	SOLFA				SACRILEGA
		LIMFA				COL.LEGA
-MFA	2	NIMFA				PLEGA
		COFA				REPLEGA
		FOFA				