La elaboración automática en los estudios hispánicos y el Banco de Datos de Prensa Española 1977 (PE77)

David Mighetto

1. Generalidades

Muchos hispanistas han tratado desde diversos puntos de vista el tema del empleo de la técnica electrónica en los estudios lingüísticos y literarios. Que yo sepa, una de las bibliografías más extensas sobre el tema se la debemos a Leopoldo Sáez-Godoy, Romanisches Seminar, der Universität Bonn (cfr. *Las computadoras en el estudio del español: bibliografía,* L. Sáez-Godoy, en «Tesavrus», XXXVIII, N.° 2, págs. 340-375. Bogotá, 1983) y a él remito gran parte de los trabajos que menciono en estas páginas.

A título de ejemplo, menciono aquí algunos trabajos de Manuel Alvar Ezquerra sobre:

— *La electrónica aplicada a la investigación lingüística,* en «Arbor», 410, págs. 15-25, 1980.

— *La asistencia de los ordenadores en la elaboración de obras lexicográficas,* en «Proyecto de Lexicografía Española», págs. 203-229. Madrid, 1976.

— *La obtención automática de índices y de sufijos,* en «Revista de Dialectología y Tradiciones Populares», XXXII, págs. 33-42. Madrid, 1976.

En el artículo *Informática y lingüística* (en «Revista Española de Lingüística», 7, 1, págs. 192-212, 1977), Manuel Alvar afirma que es imprescindible el uso de las computadoras si queremos mantener la unidad lingüística de los hispanohablantes. Unidad lingüística amenazada —al mismo tiempo que estimulada— por la influencia y la penetración arrasadora de los medios de comunicación. Con este propósito, la innovación lingüística se hace indispensable frente al arcaísmo de la lengua. En su artículo (cfr. pág. 207) dice:

> «Tendríamos que hacer inventarios léxicos y sintácticos de la prensa diaria; del contenido de los enunciados o cabeceras; del mensaje que puede comunicar un anuncio... Pero esto no puede hacerse a mano».

La propuesta citada apunta a subrayar la imperiosidad siempre creciente del hecho de no poder seguir prescindiendo de la eficacia, la rapidez y la economía de las máquinas. Su pronóstico de 1977 está hoy confirmado: la computadora ha sacado a la lingüística española de donde estaba; su aplicación ha permitido hacer grandes adelantos en lingüística española.

2. La computadora como instrumento de trabajo

La computadora es para el lingüista lo que un lápiz es para mí cuando extraigo manualmente ejemplos interesantes de un periódico: un instrumento de trabajo. Suma-

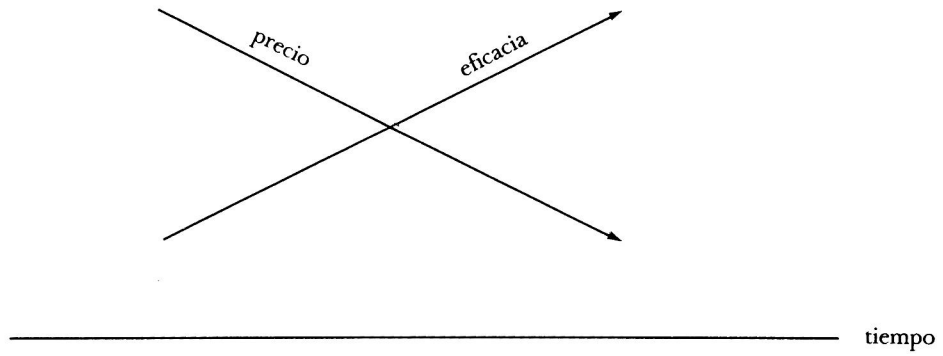
BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

mente efectivo, sí, pero nada más. Principalmente, se debe a su efectividad el hecho por el cual las computadoras merecen especial atención.

La efectividad de la computadora puede hasta llegar a ser contraproducente. Una computadora es capaz de tratar textos extensos y, p. ej., llenar miles de páginas con cifras y frecuencias, hacer listados de todas las palabras que contienen una *r* o una *s*, confeccionar listas de todas las palabras del texto que van precedidas por *de*, listar todos los sintagmas entrecomillados, hacer listas con las palabras más comunes a principios de oración, etc. En fin, todo lo que podemos formalizar en un algoritmo (instrucción o programa), la máquina puede —en principio— ejecutarlo.

Frente a la cantidad de información de salida, se corre el riesgo de ahogarse en un mar de datos. Por ello, los datos de salida que se exijan a la máquina —por un lado— tendrán que poder ser interpretables y explicables, y —por otro— tendrán que poder ser puestos en relación con otros datos de referencia. En este tipo de precaución a tomar se centra, en definitiva, el cuidado que hay que tener al usar las computadoras para fines lingüísticos.

La relación a través del tiempo entre el *precio* y la *eficacia* de las computadoras queda dada con el siguiente gráfico:



A medida que el precio ha ido decreciendo abruptamente, su eficacia ha ido aumentanto astronómicamente. Los pronósticos hechos en lo que va de los años 80 apuntan hacia un mismo tipo de tendencia para el futuro. Hoy en día no son excepcionales los sistemas de computación con capacidad de ejecutar 10 millones de operaciones por segundo, y en laboratorio existen ya sistemas que triplican esa cifra.

Fueron las mayores facilidades de adquisición las que, sin duda, representaron uno de los factores decisivos para que países como los hispanohablantes comenzaran a hacer uso de las computadoras en investigaciones de la lengua.

En nuestros países, el proceso de datos al servicio de la lingüística y la literatura españolas comenzó a tener relativa importancia en la década de los años 70. En lo que va de los años 80 se ha hecho aún más importante y constituye un proceso de informatización que indiscutiblemente está en el tiempo. Con anterioridad, los hispanistas que elaboraban automáticamente sus textos castellanos, eran investigadores localizados preferentemente en ciertos países de Europa occidental y de EE. UU. Estas naciones contaban entonces con la técnica necesaria y con el no menos imprescindible respaldo económico para la clase de actividades humanísticas en cuestión. No obstante el retraso de nuestros países, en ellos se hacen hoy estudios exhaustivos sobre el castellano con ayuda de la computadora.

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

3. Algunas investigaciones realizadas sobre soporte de computadora

A continuación señalo algunas de las áreas de investigación en las que se trabaja con apoyo de la lingüística computacional: bibliografías, geografía lingüística, lexicología contrastiva, análisis literario, enseñanza del castellano como lengua extranjera, inteligencia artificial, morfosintaxis, semántica, traducción automática.

Además de ellas, deseo señalar particularmente las siguientes investigaciones:

— en *morfología*, los trabajos de Josse De Kock.

— en *lexicografía*, el diccionario de frecuencias de Huilland y Chang-Rodríguez (1964).

— el IDE o «dictionnaire inverse» del español, cuya elaboración está a cargo de Silvia Faitelson-Weiser (1980).

— el *Diccionario Reverso* de D. Mighetto y P. Rosengren (1985).

— en *grafotaxis,* el trabajo exhaustivo *Las letras españolas: estudio grafotáctico,* de D. Mighetto y P. Rosengren (1984).

— en *fonética*, los trabajos del Departamento de Fonética del CSIC (Madrid) iniciados en 1979 y relativos a la síntesis de la voz en español y a la conversión de textos escritos a textos orales.

— en *determinación de autor* (o atribución de paternalidad literaria), Ingemar Engström, del Dep. de Lenguas Romances de Gotemburgo, está trabajando en la lengua del Quijote de Avellaneda. A partir de la edición de «Clásicos Castellanos», Engström ha establecido dos concordancias en formato KWIC del corpus de 130.000 palabras seguidas: una concordancia es alfabética inicial (6 volúmenes), y la otra está alfabetizada «a tergo», o sea, de derecha a izquierda (también 6 volúmenes). Entre los objetivos de su estudio se destaca el análisis de las características lingüísticas de la obra sobre una base de premisas de la estadística lexical. Engström ha establecido también un índice de frecuencias de las palabras gráficas.

De ese índice he sacado las siguientes cifras que cotejo con las correspondientes al recuento hecho por Ruiz-Fornells sobre *El Quijote* de Cervantes:

El Quijote de Avellaneda:		*El Quijote de Cervantes:*	
Palabra gráfica	*Frecuencia textual*	*Palabra gráfica*	*Frecuencia textual*
de	6.933	de	17.847
que	6.840	que	20.443
y	6.042	y	17.029
Quijote	1	Quijote	2.163
Quixote	750	Quixote	(?)
Sancho	785	Sancho	2.149
Dulcinea	17	Dulcinea	283

De la misma lista de frecuencias he sacado las 10 palabras gráficas (no lematizadas) más comunes en el texto de Avellaneda, a saber: *de*: 6.933; *que*: 6.840; *y*: 6.042; *la*: 3.856; *a*: 3.791; *en*: 3.171; *el*: 3.166; *con*: 2.114; *por*: 1.675; *no*: 1.654.

Esta escueta presentación de frecuencias absolutas ha sido posible hacerla gracias a la computadora y a la accesibilidad de los textos de los dos Quijotes. Por ejemplo, no lleva mucho tiempo sintetizar la información allí implícita de que, entre las pala-

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

bras más comunes, en el texto de Avellaneda, se encuentran las preposiciones, o que se mantienen las proporciones, en los dos textos, en el uso de los dos nombres *Quijote* y *Sancho*, al mismo tiempo que Cervantes tiene predilección por el de *Dulcinea*. A Engström —u otro— le tocará establecer con el tiempo si se trata de señalamientos lingüísticos capaces de determinar el autor del Quijote apócrifo.

— Entre los muchos *índices lingüísticos* se encuentran los vocabularios elaborados sobre *El Poema de Mio Cid,* el *Libro de Alexandre* y el léxico de Lope de Rueda.

— Las *concordancias lingüísticas* ocupan también un lugar destacado en las investigaciones con computadora de los últimos años: en 1976, Enrique Ruiz-Fornells editó, después de nueve años de trabajo, su primer volumen de la concordancia alfabética inicial de *El Quijote* de Cervantes, y en 1984 editó *Las concordancias del Quijote de Avellaneda.* Arriba he citado ya las otras concordancias sobre el falso Quijote. Hay concordancias del teatro completo de Valle-Inclán, de García Lorca, de la poesía de Bécquer, del *Martín Fierro,* de las obras completas de Gil Vicente, de Calderón, de *La Araucana,* de las obras poéticas de Garcilaso de la Vega, de *El Poema de Mio Cid,* de la obra de Alfonso El Sabio, del *Libro de buen amor,* de *La Celestina,* y muchas más. Mi deseo es que, con el correr del tiempo, estos trabajos estén más fácilmente disponibles de lo que lo están hoy.

— En cuanto se refiere a los trabajos realizados o en vías de realización sobre bancos de datos y bases de datos, cabe señalar —por un lado— la iniciativa de Espasa-Calpe para establecer un banco de datos iberoamericano y —por otro— los resultados alcanzados con el *Banco de Datos de Prensa Española 1977 (PE77)* a los cuales me referiré más abajo en el punto 4.

La Editorial Espasa-Calpe inició a principios de 1983 un plan de automatización por computadora con el fin de facilitar la redacción y la impresión electrónicas. Este tipo de actividades suele desembocar automáticamente en la creación de un banco de datos, o si no —según las características de disponiblidad y de acceso al material— en la institución de una base de datos. En su plan de acción, Espasa-Calpe prevé la realización de varios proyectos. Uno de ellos —a corto plazo— acomete la confección de un índice de su enciclopedia, destinado a facilitar su consulta. Otro proyecto —a largo plazo y seguramente el más interesante para la lingüística española— es la creación de un banco de datos conjunto de la comunidad lingüística iberoamericana. El día que este banco de datos esté instituido será un día de fiesta para los hispanistas y para todos los usuarios de nuestra lengua que se interesan por su estado y se preocupan por su porvenir.

4. PE77: Banco de Datos de Prensa Española 1977

Durante los últimos años muchos investigadores han señalado la importancia de estudiar el lenguaje periodístico. No faltan proyectos donde la técnica de la computadora ha facilitado el tratamiento de materiales extensos, pero la mayoría de los estudios sobre prosa periodística han sido enfocados —al menos en lo que toca al castellano— siguiendo más bien métodos tradicionales.

Con el *Proyecto PE77 (Banco de Datos de Prensa Española 1977, PE77)* alcanzamos en mayo de 1984 el objetivo de editar en la modalidad de microficha COM el texto fuente del *PE77* (con 2 millones de palabras seguidas de prosa periodística española pe-

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

ríodo junio-diciembre de 1977) y una concordancia alfabética inicial de sus palabras gráficas.

Uno de los propósitos del *PE77* ha sido el de elaborar automáticamente textos periodísticos para poder ponerlos sobre soporte de computadora con el fin de utilizarlos en la investigación lingüística y en la enseñanza de la lengua española.

El resultado de esta aspiración se halla en las 172 microfichas que ofrecemos con la concordancia lingüística en formato KWIC y el texto fuente en forma de texto seguido. Además, hemos confeccionado una lista de palabras gráficas con frecuencias de acuerdo a la fuente, a la pertenencia lingüística de autores, al tipo de artículo y a la esfera temática tratada en los artículos. Esta lista, de 1.350 páginas en 4 volúmenes, acompaña a la edición de microfichas, y la presentamos con el convencimiento de que podrá consultarse para aumentar la vista de conjunto del material y para precisar el análisis de cada una de las unidades gráficas en particular.

Con sus más de 250 millones de posiciones memorizadas, cantidad con la que se podrían llenar unas 80.000 páginas de un libro normal, la edición de las 172 microfichas constituye un material lingüístico de apreciable envergadura. Esta gran voluminosidad fue el factor determinante para que la edición se hiciese en la modalidad de microficha. El muy bajo precio de venta de la edición (cubre exclusivamente los gastos de las copias) pone el material no sólo al alcance de centros de estudio y bibliotecas, sino también dentro del alcance de investigadores y estudiantes del español en general.

Así, por ejemplo, la lista siguiente da el nombre de algunos propietarios del *PE77:* las universidades de Gotemburgo, Lund, Estocolmo, Upsala, Oslo, Helsinki, Copenhague, Católica de Lovaina, Paris-Sorbonne, Viena, Linz, Heidelberg, Salzburgo, Sevilla.

Vale indicar que cualquier forma del material del *PE77* se halla disponible solamente para fines docentes y para la investigación.

Un artículo completo con la descripción del corpus, la conformación y la codificación del corpus se suministra conjuntamente con las microfichas. No obstante, en las páginas siguientes doy una presentación escueta del *PE77.*

Con el *Proyecto PE77* hemos instituido el *Banco de Datos de Prensa Española 1977.* El texto fuente del *PE77* consta de 2 millones de palabras seguidas de prosa periodística española moderna. El 44 % del material viene del diario «El País», mientras que a la revista semanal «Triunfo» le corresponde el 56 % restante. El material, editado en España, es del período junio-diciembre de 1977.

Con *artículo* se entiende aquí alguna de las siguientes formas de expresión periodística: telegramas y télex de agencias de noticias, de corresponsales o de enviados especiales de la Redacción; crónicas, editoriales y artículos de fondo; entrevistas; relatos; reseñas.

Los artículos que conforman el *PE77* están codificados por autor (418 códigos), por fuente y, además, de acuerdo a la siguiente clasificación:

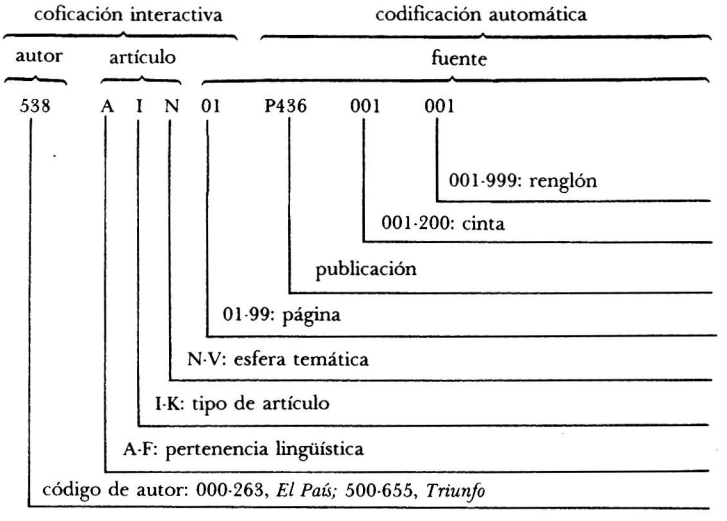
Pertenencia lingüística de autores de artículo	*Código*
Española; artículo firmado	A
Española; artículo no firmado	B
Otra lengua/traducción; artículo firmado	C
Otra lengua/traducción; artículo no firmado	D
Desconocida; artículo firmado	E
Desconocida; artículo no firmado	F

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

Tipo de artículo	*Código*
Telegrama y télex de agencia de noticias y corresponsales	I
Editorial	J
Otro tipo de artículo	K

Esfera temática	*Código*
Política y cuestiones socio-gremiales	N
Deportes; la lidia; aficiones	P
Ciencias humanas	Q
Actividades artísticas	R
Ciencias naturales/aplicaciones	S
Vida económica y comunicaciones	T
El hombre y su ambiente	U
Miscelánea	V

Resumen esquemático de la codificación del material (ejemplo)



BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes

Graford	*FT*	*El País*	*Triunfo*	*Pertenencia lingüística*						*Tipo de artículo I*				*Esfera temática*				
				A	*B*	*C*	*E*	*F*	*I*	*J*	*K*	*N*	*P*	*Q*	*R*	*S*	*T*	*U*
aconsejo	1	1	0	1							1							1
aconsejó	5	3	2	4	1						5	1	1		2			1
aconstitucionales	1	0	1	1							1	1						
acontece	5	1	4	5							5			1	2		1	1
acontecer	7	5	2	7							7	1			4	1		1
aconteceres	2	1	1	2							2	1			1			
acontecían	1	0	1		1						1	1						
acontecidos	2	1	1	1	1						2	1			1			
acontecimiento	45	24	21	34	11					1	44	12	5	4	16	1		7
acontecimientos	85	24	61	66	13		4	1	4	1	80	47		2	23		3	10
acontezca	2	2	0	2							2	1						1
acopiar	1	0	1	1							1	1						
acopio	7	2	5	5	1		1				7	1		1	2		1	2
acopla	1	0	1	1							1	1						
acoplada	1	0	1	1							1							1
acoplados	2	0	2	1							2					1		1
acoplamiento	1	0	1	1							1				1			
acoplando	1	0	1		1						1							1
acoplar	3	2	1	2	1						3	1	1					1
acoplarían	1	0	1	1							1			1				
acoplarnos	1	0	1	1							1				1			
acoplarse	4	2	2	3	1						4		2		1			1
acoplarte	1	0	1	1							1				1			
acoplen	1	0	1	1							1			1				
acopló	3	3	0	2	1						3		3					
acorazada	1	0	1	1							1				1			
acorazado	20	8	12	19	1						20		1		18		1	
acorazo	1	0	1	1							1				1			
acordaba	2	0	2	1			1				2	2						
acordaban	1	0	1	1							1							1
acordada	15	10	5	7	7			1	1		14	8			3		2	2
acordadas	1	1	0		1						1	1						
acordado	41	25	16	16	25				2		39	27			1	1	3	9
acordados	4	1	3	3	1						4	1					1	2
acordáis	1	0	1	1							1				1			
acordamos	2	2	0		2				1		1	1					1	
acordando	1	1	0	1					1			1						
acordar	9	3	6	7	2						9	2		2	2		2	1

Centro Virtual Cervantes

Graford	*FT*	*El País*	*Triunfo*	*Pertenencia lingüística*						*Tipo de artículo I*					*Esfera temática*				
				A	*B*	*C*	*E*	*F*	*I*	*J*	*K*	*N*	*P*	*Q*	*R*	*S*	*T*	*U*	
acordará	6	4	2	6					2		4	2	1	1				2	
acordarme	2	0	2	2							2				2				
acordaron	26	24	2	12	12		1	1	4		22	17	1				4	4	
acordarse	4	3	1	4							4	2	1	1					
acordasen	2	1	1	2					1		1	1						1	
acorde	23	8	15	18	5						23	4	3	1	8			7	
acordes	13	10	3	13							13	2			4	7			
acordó	52	47	5	23	28			1	3		49	34	2	1	1		5	9	
acordonado	1	1	0					1	1			1							
acordonaron	1	0	1	1							1	1							
acorrala	1	0	1	1							1	1							
acorralaba	1	0	1	1							1	1							
acorralada	1	1	0	1							1				1				
acorralado	2	2	0	2							2		2						
acorralados	6	1	5																
acorralados	5	1	4	4	1						5	2			1			2	
ACORRALADOS	1	0	1	1							1	1							
acortado	4	2	2	3	1						4	1			1			2	
acortando	1	0	1	1							1	1							

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...

Centro Virtual Cervantes

Tabla porcentual correspondiente a la producción por esfera temática

Las 2.851 cintas de artículos de *PE77* (1.966 en *El País* y 855 en *Triunfo)* contienen 1.919.391 palabras gráficas (836.721 en *El País* y 1.082.670 en *Triunfo).* Estas son las unidades que han sido presentadas en la concordancia alfabética inicial. Las materias tratadas en las cintas están caracterizadas por ocho esferas temáticas.

En la tabla siguiente se hallan los valores porcentuales de la producción por esfera temática y por publicación.

Código	*Esfera temática*	*Publicación*	*Palabras gráficas*	*% por publicación*	*% del total*
N	Política y cuestiones sociogremiales	*El País*	365.602	19,04 %	41,46 %
		Triunfo	430.223	22,41 %	
P	Deportes; la lidia; aficiones	*El País*	102.571	5,34 %	5,52 %
		Triunfo	3.459	0,18 %	
Q	Ciencias humanas	*El País*	37.031	1,93 %	3,57 %
		Triunfo	31.463	1,64 %	
R	Actividades artísticas	*El País*	82.150	4,28 %	18,84 %
		Triunfo	279.378	14,56 %	
S	Ciencias naturales/aplicadas	*El País*	10.364	0,54 %	2,67 %
		Triunfo	40.928	2,13 %	
T	Vida económica y comunicaciones	*El País*	75.086	3,91 %	5,96 %
		Triunfo	39.441	2,05 %	
U	El hombre y su ambiente	*El País*	163.500	8,52 %	21,83 %
		Triunfo	255.386	13,31 %	
V	Miscelánea	*El País*	532	0,04 %	0,16 %
		Triunfo	2.277	0,12 %	
			1.919.391	(100,00 %)	100,00 %

En lo que se refiere a la constitución del material del *PE77* por esfera temática, las cifras de esta tabla hablan por sí solas. Por ello, dejamos al lector la iniciativa de ponerlas en relación con la investigación que lleve a cabo en cada caso determinado.

BOLETÍN AEPE Nº 34-35. David MIGHETTO. La elaboración automática en los estudios hispánico...
Centro Virtual Cervantes