LEXIS, Vol. IV, Núm. 2, Diciembre de 1980

ANALISIS SINTACTICO DE TEXTOS AUTOMATIZADOS

Ramón Almela Pérez
Universidad de Murcia

I. INTRODUCCION

Las páginas que siguen tratan de mostrar empíricamente las enormes ventajas que se siguen del tratamiento computacional de textos lingüísticos para el conocimiento de esos mismos textos y del sistema lingüístico correspondiente.

El ordenador realiza cientos de miles de operaciones por segundo. Lo que es la grúa respecto al brazo del hombre, es el ordenador respecto al cerebro: un instrumento potenciador de la capacidad humana, física en el primer caso e intelectual en el segundo.

¿Cómo no acudir a un instrumento tan capacitado, como es el ordenador electrónico, cuando se trata de manejar un gran número de datos, bien porque el corpus sea muy vasto, bien porque, aun siendo pequeño el corpus, sus aspectos de estudio elegibles sean numerosos? [1]

¿Cómo dudar de que con el ordenador se pueden reunir inventarios mucho más completos y accesibles[2] que los que un cerebro humano sería capaz de hacer y retener[3]?

Tal vez alguien tenga el siguiente escrúpulo: ¿es posible estudiar la poesía a base de números? La poesía en sí misma no puede ser explicada, ni con números ni sin números. "La poesía se explica sola; si no, no se explica. Todo comentario a una poesía se refiere a elementos circundantes..."[4]

Pero, puesto que "toda poesía se nos presenta en forma de lenguaje, en construcciones verbales"[5], "nada se opone a priori en el hecho poético mismo a una tentativa de observación y de descripción científica"[6]. Si el "alma" de la poesía es inanalizable, su "cuerpo" es susceptible de un estudio tan objetivo como el que se pueda realizar sobre los demás acontecimientos[7]

---

1 Miller, G.A. (1973). *Langage et communication.* Paris. Ed. C.E.P.L. Pag. 95
2 Rodríguez Adrados, F. (1976). "Utilización de ordenadores en problemas de lingüística". *Revista de la Universidad Complutense,* nº 102, vol. XXV, marzo-abril, pag. 7
3 Tournier, M. (1975). *Un vocabulaire ouvrier en 1848. Essai de lexicométrie.* Paris. Ed. Ecole Normale Superieure de Saint-Cloud et C.N.R.S. (Policopiado). Pag. 310.
4 Son palabras de Pedro Salinas, que recoge Gerardo Diego (1970) en *Poesía española contemporánea* (Antología) Madrid. Ed. Taurus. Pág. 303.
5 Vossler, K. (1960). *Formas poéticas de los pueblos románicos.* Buenos Aires. Ed. Losada, Pag. 15.
6 Cohen, J. (1974). *Estructura del lenguaje poético.* Madrid. Ed. Gredos. Pag. 25.
7 Hörmann, H. (1973). *Psicología del lenguaje.* Madrid. Ed. Gredos. Pag. 121.

Ahora bien, dados el desarrollo actual de la técnica computacional y la efervescencia lingüística, creemos que es más rentable y fecundo que el lingüista y el programador sean, por regla general, personas distintas.

No es que a priori esté vedado al lingüista adentrarse en la programación, no. Lo que afirmamos es que al lingüista no debe preocuparle la programación, porque esta no es de su competencia específica[8]. Dedicándose a ambas tareas, la eficacia en una y otra quedaría lógicamente debilitada.

Sin embargo la colaboración entre el lingüista y el programador es imprescindible[9]. El diálogo entre el lingüista y el programador, o mejor, entre las funciones de investigación lingüística y de programación, no se interrumpe hasta el final, teniendo dicho diálogo la misión sea de resolver dificultades, sea de comprobar la marcha o los resultados[10]

Los ámbitos lingüísticos en los que puede penetrar la Informática son, de una o de otra forma, todos. El tratamiento de la morfosintaxis, por ejemplo, requiere la lematización. El ordenador no sabe si CORRER es verbo, si NOS es pronombre, si MESA es sustantivo

Por ello hablaré en primer lugar de la lematización, tal y como yo la he experimentado, y luego pondré un ejemplo referido a la morfosintaxis.

II. PROCESO LEMATIZADOR

El trabajo, tal como lo voy a exponer, lo hice como tesis doctoral, y versó sobre las poesías completas de Pedro Salinas. La parte técnica corrió a cargo de los miembros del Instituto de Lingüística Computacional, de Pisa (Italia), dirigido por el Prof. Zampolli[11].

El texto que seguí fue la 2a. edición de *Poesías completas* de Pedro Salinas, preparada por la hija del poeta, Solita Salinas de Marichal, y publicada por Barral Editores en 1975.

La lematización (Devoto habla de "lemación") consiste en conducir toda forma lingüística a su matriz y describir su categoría, o sea, en acoplar cada "ocurrencia" y "vocablo" a su respectivo "lema". Así, en un hipotético texto, las diez veces que aparece el "vocablo" o "forma" YENDO son otras tantas

---

8 Kock, Josse de (1974). *Introducción a la lingüística automática en las lenguas románicas.* Madrid. Ed. Gredos, Págs. 27-28.

9 Muller, Charles (1977). *Principes et méthodes de statistique lexicale.* Paris. Ed. Hachette Université.

10 Tournier, M. (1967). "Vocabulaire politique et inventaires sur machine". *Cahiers de Lexicologie,* 10, 1, pp. 67-81.

11 Pedro Salinas (1891-1951), español, Profesor universitario, enseñó desde 1935 en universidades de Estados Unidos, y está enterrado, por voluntad propia, en Puerto Rico.

"ocurrencias" o "palabras" que corresponden al "lema" IR. En estos conceptos sigo a Ch. Muller.

Las siglas que empleo más abajo (SEG, RAZ, CONF) se refieren a otros tantos libros poéticos en que se agrupan los poemas de Pedro Salinas. Dichos libros son trece:

PRE. . .Presagios
SEG. . .Seguro azar
FAB. . .Fábula y signo
VOZ. . .La voz a tí debida
RAZ. . . Razón de amor
LAR. . . Largo lamento
CONT. . El Contemplado
TO. . . Todo más claro
CONF. . Confianza
PI. . . . . Poemas inéditos
PP. . . . Primeras poesías
PS. . . . . Poemas sueltos
PU. . . .. Poesía última

A) PREEDICION

El texto fue introducido en el ordenador tal como viene en la edición que manejo, con los límites aclarados en el apartado dedicado a "El autor y la obra". Solamente hubo una clase de adición sustantiva y sistemática al contenido original: la identificación de pertenencia y de numeración de todo poema y de todo título. Cuanto el ordenador nos ha "dicho" después no lo "sabía" desde el principio, sino que se le suministró en una fase posterior. Pondré 3 ejemplos del único contenido añadido inicialmente al texto escueto.

El poema no 3 de SEG figura así:

3

OTRA TU

"No veo la mirada
. . . . . "

y se le dio al ordenador así:

o/o SEG, 3 T OTRA TU o/o SEG, 3 No veo. . .

El poema no 43 de RAZ no trae en la edición ni título ni número; comienza la composición directamente con el verso primero: "No te guar-

des. . .". Se le dio al ordenador así:

o/o RAZ, 43 No te guardes. . .

El poema no 9 de CONF está encabezado por el título, sin número:

PRESENTE SIMPLE
"Ni recuerdos. . .
. . . . . ."

y fue introducido así:

o/o CONF, 9 T PRESENTE SIMPLE o/o CONF, 9 Ni recuerdos. . .

Con excepción de estas innovaciones, necesarias desde el punto de vista técnico, no corregí nada, salvo algunos leves errores de imprenta, el texto editado.

B) DESCRIPCION GLOBAL

El texto, tal como acaba de ser descrito, se perforó en unas 20,000 fichas aproximadamente. Estas, perforadas, se introdujeron en el ordenador precedidas de las que contenían el programa. Como fruto tuve la Lista Texto primera (que denominaremos LT 1ª, así como a toda Lista Texto la llamaremos LT), que era igual al texto editado; es decir, contenía las mismas indicaciones del original y de la misma forma: mayúsculas y minúsculas, indicaciones de las páginas del texto original, acentos, etc.

Naturalmente había errores, pero estos no estaban programados. Lo que sí estaba programada era la corrección de los mismos. Los detecté comparando la LT 1ª con el original. ¿De qué clase eran los errores? Pues desde la existencia de un punto en lugar de una coma, p. ej., o la omisión de una palabra, hasta la repetición de un mismo verso.

Se listó una 2ª LT, con menos errores ya; se volvieron a corregir; así se siguió este proceso hasta obtener una LT que era la reproducción fiel del texto original preeditado.

A partir de la última LT se obtuvieron tres índices y unas concordancias. Los tres índices lo eran de formas, de vocablos: *amores, amor, buscabas, me, tenues, ya, . . .* eran otras tantas formas. Pues bien, de éstas se obtuvieron:

– lista alfabética directa, más la frecuencia
– lista alfabética inversa
– lista decreciente de frecuencias.

Las concordancias por formas eran un listado de casi 6,000 páginas. Su estructura era la siguiente: por orden alfabético cada forma aparecía encabezando un grupo de tantas líneas como número de veces ocurría en el texto original, y en cada línea la palabra "pivot" (destacada por dos asteriscos a su derecha y otros dos a su izquierda) era precisamente esa, la que servía de cabeza –su nombre propio es "exponente"– en cada caso, pero venía acompañada, a uno y otro lado, de un contexto de unas veinte palabras como término medio.

Ya este era un gran producto, matriz de otros ulteriores. Y hasta aquí la fase de despojo del texto. Vino después la fase de lematización.

Sobre el texto así preparado, fabriqué el sistema de lematización que ofrezco en otra parte de la tesis. Aplicando este sistema, las concordancias por formas se transformaron en concordancias por lemas. Junto a cada forma (el número de formas era de casi 11,000) colocaba el lema a que pertenecía y la información correspondiente; p. ej., *bueno* es el lema de las formas *bueno* y *buenos*, *sacar* es el lema de *sacabas* y *saqué*, y así otros. La información pertinente de cada lema venía dada por las letras según la posición que, por convención sistemática, les tenía asignada.

Como cada forma podía pertenecer, si tenía contextos diversos y gozaba de una, aunque fuera leve, polisintaxia, a más de un lema, y aun en el caso de que no se diera ninguna de esas circunstancias, pues cada contexto podía presentar algún aspecto "sui generis", era necesario leer todas las líneas, todos los contextos, a fin de evitar en lo posible errores en la asignación del lema o de los lemas apropiados.

Siendo más de 85,000 los contextos y conteniendo cada uno unas 20 palabras aproximadamente, el intento de una correcta lematización me obligó a la lectura de 1.700.000 ocurrencias; lectura que tuve que repetir en el control de la lematización, acompañada, las dos veces, del análisis lingüístico.

Esta operación, la lematización, fue el eje de la posterior archivación de datos en el ordenador. La información que se le comunicó en esta fase es toda la que, no estando en el original, conoce el ordenador.

Cualquier programa, objetivo, índice, listado etc., puede contar sólo con las informaciones que les proporcionaron las fichas perforadas en las dos etapas: la primera, o fase de despojo, que le ofrecía el texto del poeta, y la segunda, o fase de lematización, que le añadía el análisis del investigador.

La lematización se vertió en unas 40,000 fichas aproximadamente. Se imprimió un listado primero de concordancias por lemas; lo corregí, para lo cual releí los 85,000 contextos; salió un segundo listado, que volví a corregir; así se prosiguió hasta obtener la lista definitiva.

Las fichas perforadas, una vez leídas por el ordenador se vertieron en discos magnéticos, que son los que se conservan y utilizan como fuentes de

datos. En total son 20 los discos magnéticos que, para el conjunto de mi tesis, han sido utilizados.

Con la información lemática se construyeron varios índices y unas concordancias. Estas tenían una cabeza general: el lema, exponente, colocado alfabéticamente; y varias cabezas parciales: tantas como formas distintas están englobadas dentro de cada lema. Bajo cada forma aparecían todas sus ocurrencias en contexto, o sea, tantas líneas como veces aparece una forma; así en el caso, p. ej., del lema *ir*, aparecían las formas *iba, irías, vamos, yendo,* etc., cada una con sus respectivos contextos.

Los índices fueron estos:

–lista alfabética de lemas

–lista decreciente de frecuencias de los lemas

–lista de secuencias lemáticas

–lista de frecuencias de informaciones gramaticales

–lista de lemas por orden de categorías gramaticales y suborden de libros poéticos

–lista de lemas por orden de libros poéticos y suborden de categorías gramaticales.

El resto de índices, listados, estadísticas, etc., de tipo fonológico, léxico, métrico, etc., se obtiene aplicando a la materia almacenada en la memoria del ordenador, los diagramas y reglas preparados por mí, lingüista, y convertidos en programas por un informático.

Dichos listados posteriores, como en el caso de los índices y concordancias emanadas de la LT, no precisan corrección, como tal, sino una mera comprobación, un sondeo, de que el programa funciona; tarea que, en el caso de programas ya empleados, es más rutinaria que minuciosa.

¿Por qué no se precisa corrección? Porque el ordenador sabe que 1 + 1 = 2; es decir, porque el ordenador no se equivoca calculando, porque el cálculo radical que realiza consiste en una suma, y porque esos resultados son el último fruto, gráfico y ordenado, según convenciones, de un cálculo.

Hasta aquí la descripción global de la preparación y elaboración lingüísticas en el proceso técnico. Nada digo de la dimensión informática: mecánica, programadora, analista, . . .; esta es la otra cara, imprescindible pero diversa, de un trabajo como el presente.

III. UN EJEMPLO

Entre los numerosos resultados que he obtenido de la tesis (más de veinte tomos), mostraré a continuación uno, breve, referente a emparejamientos sintácticos.

1º. ATENDIENDO AL VERBO

A. PRONOMBRE PERSONAL MAS VERBO

Haciendo dos cortes horizontales tendríamos, junto a otras posibles, estas dos agrupaciones:

1ª: FUNCIONAL

	Cantidad	*o/o*
Enclíticos (me, te, se, nos, os, la, las, le, les, lo, los).....	3540	90'74
Sujetos (yo, tú, nosotros, nosotras, vosotros, vosotras, usted).....	247	6'33
Otros (mí, ti, sí, conmigo, contigo, consigo, él, ellos, ella, ellas, ello).....	114	2'92
	3901	

2ª: PERSONAL

	Cantidad	*o/o*
1ª Persona	890	22'81
2ª persona	873	22'37
3ª persona	2138	54'80
	3901	

Las cifras resultantes no tienen cimas sorprendentes. Los enclíticos

confirman su adhesión sistemática al verbo. Sin embargo, de los pronombres sujeto se podía esperar una mayor proximidad al verbo: ¿busca con ello Salinas la consecución de un ritmo sintáctico más lento, más saboreado?

Creo que se debe subrayar que el 65'92o/o de las ocurrencias de los pronombres personales van seguidos de alguna forma verbal.

Para terminar, dos curiosidades. El presente de indicativo (50'67o/o del total del verbo) es el preferido de los pronombres personales: de los 3901 que preceden a verbo, 2500 están seguidos de ese tiempo y modo. Por otro lado, *ti* acaba frase casi en la mitad de sus apariciones: 130 / 274.

B. VERBO MAS OTRA CATEGORIA

(Hemos elegido las que pueden ser más interesantes).

Categoría	*Cantidad*	*o/o de ocu. cat.*
Adjetivo	616	8'39
Adverbio	1201	18'07
Artículo	1418	19'79
Contracto	196	15'87
Cuantificador	518	19'22
Preposición	2282	21'95
Pron personal	304	5'13
Sustantivo	809	4'36
Verbo	906	6'60

La categoría menos "buscada" por el verbo es el sutantivo: esto es llamativo. No sorprenden las abundantes presencias de preposición y artículo. El adverbio es más esperado de lo que de hecho resulta en la realidad; su adhesión posicional al verbo resulta baja.

2o. ATENDIENDO A LA PREPOSICION

A. SECUENCIAS

¿Qué categorías siguen a las preposiciones?
Veamos algunas de las más significativas.

Categoría	*Cantidad*	*o/o*
Adjetivo	259	2'49
Artículo	2770	26'64
Cuantificador	792	7'61
Pron. no personal	376	3'61
Pron. personal	653	6'28
Sustantivo	2532	24'36
Verbo	927	8'91
Otras	2085	20'05
	10394	

Como única observación particular diría que las cantidades del verbo y del pronombre personal son altas respecto a lo esperable como término medio. En el primer caso no es tan corriente en el sistema que el verbo siga a la preposición; en el segundo, la causa es la alta frecuencia del pronombre personal en nuestro poeta.

B. DISTRIBUCION DE PREPOSICIONES EN CATEGORIAS

(Van entre paréntesis los o/o respecto de las ocurrencias totales de cada preposición)

Categoría	*A*	*CON*	*DE*	*EN*	*PARA*	*POR*
Adjetivo	10 (0'06)	26 (3'89)	111 (3'18)	74 (3'33)	1 (0'34)	26 (2'79)
Artículo	375 (25'37)	196 (29'34)	679 (19'47)	892 (40'14)	33 (11'41)	318 (34'15)
Pronombre	230 (15'56)	81 (12'12)	225 (6'44)	271 (12'19)	42 (14'53)	70 (7'51)
Sustanti.	137 (9'26)	156 (23'35)	1277 (36'59)	387 (17'41)	7 (2'42)	113 (12'13)
Verbo	306 (20'70)	6 (0'89)	259 (7'42)	34 (1'53)	119 (41'17)	57 (6'12)

C. COMENTARIO

En términos absolutos la preposición más irregularmente distribuída es *de*, y la menos es *para*: ¿quizá porque son la más y la menos abundante, respectivamente entre estas seis?

En cifras relativas, sin embargo, la más irregular es *para*, y las menos irregulares son *por* (con 31'36 de oscilación entre los o/o mayor y menor), y *de* (con 33'41 de oscilación).

Por categorías son dignas de mención las adhesiones de *para* al verbo, de *en* y *por* al artículo, y de *de* al sustantivo. Los desafectos que descuellan son los de *a* y *para* hacia el adjetivo, los de *con* y *en* hacia el verbo y el de *para* de nuevo, hacia el sustantivo.

En conjunto, las categorías, atendiendo a la cantidad de porcentajesque en ellas depositan las diversas preposiciones, se jerarquizan así:

Artículo	159'86	(suma de los o/o parciales)
Sustantivo	101'16	(id.)
Verbo	77'83	(id.)
Pronombre	68'35	(id.)
Adjetivo	13'59	(id.)

Lo único llamativo es, a mi juicio, la alta cantidad relativa del verbo: se la proporcionan *a* y *para*.

Comparando los comportamientos de *a, de* y *en* (las preposiciones más numerosas pues con sus 7190 apariciones totalizan el 69'17o/o de las ocurrencias de las preposiciones) respecto al adjetivo, sustantivo y verbo, constatamos que:

—son semejantes respecto al adjetivo
(0'6; 3'18; 3'33)
—*de* se inclina por el sustantivo
(9'26; *36'59;* 17'41)
—*a* se inclina por el verbo
(*20'70;* 7'42; 1'53)
—*en* casi ignora al verbo
(20'70; 7'42; *1'53)*

3º. ATENDIENDO AL ARTICULO

Cantidades de las secuencias de artículo más las categorías indicadas. Los porcentajes se refieren a la cantidad total de ocurrencias del artículo.

	Adj. Calif.	*Pronombre*	*Sustantivo*	*Verbo*
El	155	51	1733 (82'09º/o)	51 (2'90º/o)
LA	224	91	1826 (80'93º/o)	1
LO	172 (30'12º/o)	353 (61'82º/o)	4 (0'70º/o)	3
LOS	56	38	940 (82'23º/o)	3
LAS	21 (1'94º/o)	20	867 (80'20º/o)	3

En cifras absolutas poco se podría deducir del cuadro anterior. Por eso recurriré a los procentajes. El artículo de más irregular distribución es *lo*: el 61'82º/o preceden a un pronombre, en concreto al *que* enunciativo no correlativo, o sea la secuencia *lo que;* no se encuentra una similar respuesta para el 30'12º/o, que es elevado de su precedencia ante adjetivo, lo cual puede hablar de elección de estilo.

Prácticamente inexistente (8'06º/o) es el acompañamiento que ofrece al resto de las categorías.

Por el lado opuesto, *las* sólo en dos de cada cien ocurrencias va seguida de adjetivo.

Salvo en el caso de *lo*, el artículo precede mayoritariamente al sustantivo. Es curioso observar el equilibro de estas cantidades: oscilan entre 80'20º/o y 82'23º/o.

Sobresale la frecuencia de *el* más verbo; más de la mitad de esas ocurrencias (32 sobre 51) de verbo son infinitivos.

4º. ATENDIENDO AL CUANTIFICADOR NUMERAL

No se puede afirmar que toda palabra precedida de cuantificador vaya, ipso facto, cuantificada; pero del uso más o menos frecuente de una secuencia cuantificador + otra categoría se puede entrever la dirección cuantificadora.

Del cuadro que sigue destacaríamos el 5'47º/o del doble cuantificador, el casi 60º/o del sustantivo, aunque este dato no es demasiado extraño, y el mínimo 3'35º/o del verbo, que, no por pequeño, es insignificante. He aquí cifras de las categorías que siguen al cuantificador:

Categorías	*Cantidad*	*º/o*
Actualizador y adjetivo	28	4'94
Adverbio	9	1'59
Artículo, conjunción, contracto, locución y preposición	126	22'26
Cuantificador	31	5'47
Pronombre	25	4'41
Sustantivo	328	57'95
Verbo	19	3'35
	566	

5º. ATENDIENDO A LA SECUENCIA ADJETIVO–SUSTANTIVO

Para Salinas los adjetivos tienen bastante autonomía sintáctica respecto de los sustantivos. La mitad de los adjetivos no van unidos inmediatamente a los nombres.

Por otra parte, también los sustantivos son elocuentes de por sí poéticamente: su noción, su posición, su función, . . . aunque les falte su muy habitual acompañante. Sólo dos de cada diez ocurrencias de los sustantivos van junto a un adjetivo.

Salinas muestra predilección por los adjetivos pospuestos; para mí puede ser esto un síntoma más de la desnudez y de la importancia posicionales y conceptuales con que usa los sustantivos.

Veamos las cifras:

		o/o	
Secuencia	*Cantidad*	*Total secuen;*	*ocu. adje.*
Adje. + Sust.	1261	36'17	17'18
Sust. + Adje.	2225	63'82	30'32
	3486	47'50	