Producción automática de abstractos.*

por Ricardo Salvador Urbikova

* El autor da las gracias al *Chemical Abstracts*, a la Ohio State University y, especialmente, a sus dos compañeros de proyecto, J. Rush y A. Zamora, sin cuya colaboración este trabajo no hubiera sido posible.

A pesar de recientes adelantos en la tecnología de la documentación, la producción de abstractos de alta calidad todavía tiene que hacerse manualmente. Por consiguiente, la preparación de abstractos y sus índices vienen a representar más de la mitad de los gastos en tiempo y dinero por parte de las editoriales de revistas de abstractos. Así, pues, uno tiene que tener en cuenta factores de costo y tiempo cuando quiera comparar la producción de abstractos por medios manuales y automáticos.

Evidentemente existen dos alternativas como posibles soluciones al problema de la producción manual de abstractos:

1. Disponer de que los abstractos sean preparados por los propios autores como prerrequisitos a la publicación de sus artículos.
2. Mecanizar la producción de abstractos.

En efecto; los editores de las revistas han tratado estos últimos años de convencer a los autores de incluir los abstractos. El éxito, sin embargo, ha sido muy variable, además de que estos abstractos han demostrado ser de un valor cuestionable.

Por otro lado, la escasez de abstractores profesionales y factores de tiempo, costo y calidad han dado impulso hacia la investigación de métodos para producir abstractos e índice automáticamente. Así, pues, vamos a describir una técnica nueva para producir automáticamente abstractos de lo llamado «full text», es decir, a partir de documentos enteros.

Primero repasemos brevemente las funciones de un abstracto, que son las siguientes:

1. *Anunciamiento.*—Para dar alerta sobre la existencia de un documento.
2. *Filtrado.*—Para juzgar la pertinencia de un original.
3. *Sustitución.*—Para actuar como fuente de información en vez del original.
4. *Búsqueda retrospectiva.*—Para buscar uno entre una colección de documentos.

Asimismo se pueden agrupar los abstractos en cuatro clases:

1. *Indicativo.*—Dando al lector la suficiente información como para decidir el mismo si le interesa o no leer el documento original.
2. *Informativo.*—Procurando la suficiente cantidad de información para que el abstracto sirva en vez del original.
3. *Crítico.*—Una crítica del documento original, pero escrita en forma de abstracto.
4. *Extractos.*—Dando excertos de frases extraídas directamente del documento original.

Se puede también enfocar un abstracto con arreglo a su orientación, es decir, en función a su aplicación o utilización particular. Este es un factor importante y que hemos tenido en cuenta en nuestro sistema porque permite al usuario decidir él mismo lo que desee o no en un abstracto.

Además, los abstractos que producimos automáticamente se caracterizan por lo siguiente:

1. Una dimensión aproximada del 10 por 100 del original. (Este corte es arbitrario porque evitamos un criterio estricto de límites.)
2. Se utiliza en el abstracto la misma terminología del original, es decir, producimos extractos con la misma fraseología escrita por el autor.
3. Con excepción de los resultados finales, se procura excluir en el extracto datos numéricos, expresiones cardinales o sentencias de un valor indefinido.
4. Se evita incluir en el extracto observaciones preliminares, citaciones, notas al pie de página, expresiones parentéticas, referencias o resultados de otras personas o trabajos, datos catalogados o históricos, ecuaciones, tablas, esquemas, figuras, etc.
5. Se excluyen datos negativos, a menos de que se saquen de las conclusiones del documento.
6. Se excluyen metodología de experimentos, montaje de aparatos, preparación de muestras, etc.
7. Se excluyen ejemplos, explicaciones, opiniones, comentarios, comparaciones, en general declaraciones de tipo especulativo o subjetivo.

Por otro lado, a esta lista de exclusiones hay que añadirle factores positivos y deseables dentro de un abstracto:

1. Incluir el objetivo o propósito del proyecto, investigación o tratado.
2. Incluir los resultados y conclusiones.

En resumidas cuentas, que para producir abstractos de esta clase hay que contar con dispositivos especiales para primero poder identificar y luego eliminar o aceptar frases de documentos.

La mayoría de los sistemas desarrollados con anterioridad al nuestro se han basado equivocadamente en métodos estadísticos a partir de la frecuencia de las palabras como criterio único para seleccionar o rechazar frases del documento original. Estos métodos parecen contrarrestar el esfuerzo intelectual empleado tradicionalmente en producir abstractos, y se ha demostrado que no producen abstractos de calidad alguna.

Nuestro sistema, en cambio, se basa en métodos de *sugestión, colocación* y *titulación.* Estos tres métodos pueden resumirse diciendo que utilizan una técnica de inferencia contextual por donde con palabras o conceptos claves y con ayuda de su contexto circundante se puede deducir si una frase del documento debe o no incorporarse al abstracto. Describiremos cada uno de estos tres métodos, pero antes mencionaremos que nuestra unidad básica del trabajo es el artículo entero o por lo menos el párrafo, ya que aquellos sistemas que utilizan solamente la frase o sentencia como unidad de trabajo son inadecuados porque dentro de todo documento existe una cierta interdependencia o referencias intercruzadas entre sus elementos, que son las sentencias y el párrafo con resto del artículo.

METODO DE COLOCACION

En este método aprovechamos el orden y colocación de ciertos elementos lingüísticos de una sentencia para descubrir su valor semántico con relación al resto de la frase. En nuestro caso la primera cláusula es primordial para la identificación del significado semántico de la sentencia entera.

Por ejemplo, la frase:

> «Evidentemente esto no está bien hecho, porque uno piensa en seguida que ...»

Evidentemente empieza la frase, con lo cual se deduce que lo que continúa es claro, y si una cosa es evidente y clara, no es importante y no pertenece al abstracto. Por otro lado, *esto* es una anáfora colocada en la primera cláusula y referencia intercruzada a la sentencia anterior. Si deducimos que la frase mencionada arriba es de rechazarse, habrá que excluir la anterior a ella por motivo de estar las dos interrelacionadas directamente por medio del pronombre *esto.*

Así, pues, si descubrimos que la primera cláusula es negativa, como el caso anterior, y si dentro de la sentencia no encontramos algo más positivo, esta primera cláusula se elimina, pero acarrea que automáticamente podamos eliminar el resto de la sentencia, porque una sentencia sin su primera cláusula no tiene sentido.

En cambio está permitido eliminar otras cláusulas que no sean la primera y obtener aún resultados sensatos.

Por ejemplo:

> «La casa era muy bonita en el invierno, aunque era más confortable en el verano.»

Sentencia que puede truncarse a:

> «La casa era muy bonita en el invierno.»

Dentro de las cláusulas está permitido también eliminar expresiones preposicionales:

«La casa era muy bonita.»

Lo cual conserva el sentido de lo original, pero pierde en cuanto a su calificación.

El método de colocación lingüístico naturalmente depende del estilo del escritor. Para ello nuestro programa tiene incorporado un analizador sintáctico parcial que descubre la estructura superficial de la sentencia, lo cual es suficientemente satisfactorio para poder operar gramaticalmente con el texto.

METODO DE SUGESTION

Dijimos anteriormente que no queríamos incluir en los abstractos opiniones, nociones subjetivas o expresiones cardinales. Bajo este criterio entonces podríamos eliminar la frase «la casa era muy bonita», porque el concepto «muy bonita» encierra una noción cardenal y subjetiva.

En general podemos decir que contamos con unas 2.000 palabras claves, expresiones guías e idiomáticas y clisés que proveen una indicación inequívoca sobre opiniones, comentarios, observaciones, etc. Estos conceptos que tenemos en diccionarios en discos contienen un atributo semántico para ser emparejados con elementos lingüísticos existentes en las sentencias del documento. Existen, por ejemplo, palabras guías que sabemos que a priori predisponen al lector

de que algo importante va a decirse (y naturalmente desearíamos incluir en el abstracto). Conceptos claves y positivos como «nuestro trabajo», «este proyecto», «la presente investigación», «importantísimo», son expresiones que satisfacen este criterio.

Por otro lado captamos opiniones, explicaciones, comentarios, etc., con palabras guías negativas como «por ejemplo», «porque», «claramente», «creemos», «figura 6», «previamente», «sin embargo», «pero», interrogaciones, comillas, exclamaciones...

El peso o gravamen de estas palabras guías también depende de su posición dentro de la sentencia. Por ejemplo, frases que empiezan con «un» «algún», «muchos», es probable que incluyen descripciones más indefinidas que aquellas frases que contengan este artículo o adjetivos en situación más central. La razón de ello es que estas palabras tienen una función muy cuantitativa cuando empiezan una frase.

Lo mismo ocurre con sentencias que empiezan con participios o preposiciones, los cuales tienden a dar a la frase un sentido condicional indicando suposiciones, conjeturas o explicaciones.

> «Dado que en Madrid hay muchos coches, ocurren accidentes» (conjetura).
> «Algunos investigadores intentaron incluir los rasgos fundamentales de la personalidad...» (comentario).
> «En otro experimento se utilizaron técnicas similares...» (indefinido).

METODO DE TITULACION

Este método se basa en un glosario de palabras del título y subtítulo (excluyendo palabras nulas y funcionales).

Su lógica se basa en que el mismo autor en el título ha descrito en pocas palabras la esencia del trabajo, lo que aprovechamos para asignar automáticamente un positivo significado semántico substantivo a las palabras del título.

Sentencias que contienen palabras substantivas que también aparezcan en el título tendrán algo más peso que aquellas que no las contengan (asumiendo igualdad de condiciones en otros aspectos).

Mencionamos anteriormente que había que definir un abstracto no solamente en función de sus datos internos, sino también en función de su utilización por el usuario. Esto convierte nuestro sistema en uno de recuperación de datos.

Utilizando el mismo mecanismo de la titulación, podemos «desde fuera» incluir en el diccionario de palabras guías un número de conceptos substantivos que le interesa al usuario y darles un valor positivo para que sentencias que las contengan puedan tener una mayor oportunidad de salir en el abstracto. Por ejemplo, se podría analizar los archivos del Instituto Nacional de Previsión y construir abstractos desde el punto de vista Seguridad Social incluyendo en el diccionario conceptos positivos como asistencia sanitaria, hospitalización, maternidad, invalidez, vejez, accidentes de trabajo, desempleo, etc.

REFERENCIAS INTERCRUZADAS DE FRASES

Estas referencias dan mucha información sobre la relación y estructura lógica del texto.

Hemos visto cómo la primera cláusula de la sentencia es indispensable para el entendimiento de la frase entera. Estas primeras cláusulas pueden contener palabras que son de referencia intercruzadas de frases. Palabras como «este», «esto», «su», «tales» son ejemplos de referencias que indican que la frase que los contiene requiere otra como antecedente semántico.

Por otro lado, existen referencias intercruzadas de otro tipo. En vez de utilizar el pronombre, utilizan el mismo nombre de la otra frase. Este método de anáforas sirve para detectar frases relacionadas entre sí, llamémoslo variaciones sobre el mismo tema, porque si descubrimos que una de ellas tiene que eliminarse, la otra probablemente que también, y viceversa.

CRITERIOS DE FRECUENCIA

Mencionamos más arriba que criterios exclusivamente basados en frecuencias son nocivos para la extracción de frases. En efecto, no se puede producir abstractos aplicando la ley de que una frase es importante solamente porque proporcionalmente la frecuencia relativa de su contenido en conceptos es grande.

Nosotros, y siguiendo la regla de la teoría de la información, hacemos lo contrario y aplicamos la ley de la frecuencia inversamente modificando dinámicamente el peso semántico de las palabras guías según su frecuencia en el documento original.

Si cierta palabra guía aparece en el original un número de veces superior a un valor umbral (que depende de la extensión del texto entero), se hace una de dos: o se baja el valor semántico si era positivo o se reduce el peso negativo a menos negativo.

DICCIONARIO

El diccionario que llamamos *World Control List* (listado de palabras de control) consiste en un grupo de palabras guías clasificadas alfabéticamente. Cada elemento de este listado contiene además de la palabra o expresión dos argumentos: un valor semántico y/o un atributo sintáctico, de la manera siguiente:

Serie de letras, peso semántico, atributo sintáctico.

Los códigos semánticos utilizados para representar un valor significativo se describen a continuación y están clasificados en orden jerárquico de implementación dentro del programa:

- M Valor negativo máximo, equivalente al NOT del álgebra booliana.
- I Valor positivo máximo, algo que inequivocadamente descubre algo importante («esta investigación», «importantísimo»).
- A Muy negativo, descubre frases que no queremos en un abstracto («claro», «posibilidad»).
- K Valor positivo menor que I («a nuestro juicio», «digno de atención»).
- B Valor negativo menor que A («sin embargo», «por otra parte», «no obstante»).
- E Para intensificadores y determinantes, valor negativo dependiente de su colocación en la frase («muchos», «mas», «muy», «pesado», «alto»).
- L Calificativos de introducción. Valor negativo dependiente de colocación («una vez», «se», «participios»).
- C Para referencias intercruzadas («esto», «tales»).
- H Términos de introducción a frases modificadoras («cuyos», «es decir»).
- G Asignado por el programa a conceptos del texto que aparecen en el título.
- D Palabras nulas.
- J Continuación del código semántico de la palabra anterior.

Los atributos empleados por el analizador sintáctico son los siguientes:

- A Artículo.
- C Conjunción.
- N Pronombre.
- P Preposición.
- O Exclusivo de OF.
- Q Exclusivo de TO.
- R Exclusivo de AS.
- V Verbo.
- W Verbo auxiliar.
- X Exclusivo de IS, ARE, WAS, WERE.
- Z Negativos («ningún», «no»).

Ejemplos del diccionario *WCL:*

A nuestro juicio*K
A pesar de*B
A* *P
Ahora*A
Con* *P
Descrito*A
Digno de atención*K
Esta investigación*I
Esto es*B

Evidente*A
Importante*K
Más bien*A
Naturalmente*A
Ningún*E*Z
No obstante*B
No*L*Z
Nuestro*K*N
Otros*E
Por otra parte*B
Posibilidad*A
Principal*K
Siguiente párrafo*A
Son* *X
Tabla*A
Vamos describir*I
Fig. 1

IMPLEMENTACION DEL PROGRAMA

La figura 2 muestra de una manera general el esquema del sistema.

El programa central MAIN llama como una docena de módulos, cada uno de ellos, a su vez, llamado una serie de subrutinas, y tiene la función de coordinar, correlacionar y controlar los pasos básicos para la implementación del sistema.

Los datos de entrada, es decir, el artículo entero, se almacena en memoria en un área de entrada-salida (I/O AREA), y a partir de ella se crea una TABLA conteniendo los datos del I/O AREA organizados de tal manera que permita un proceso eficiente en orden alfabético, cronológico y jerárquico de los datos del I/O AREA.

Esta tabla contiene por cada palabra del texto los siguientes elementos en un campo fijo de ocho caracteres, distribuidos de la siguiente manera:

1 carácter - la longitud de la palabra
3 caracteres - dirección de la palabra en en el I/O AREA.
1 carácter - valor semántico.
1 carácter - atributo sintáctico.
2 caracteres - vector alfabético.

En la tabla, el elemento n-simo corresponde a la palabra n-sima del I/O AREA. Sin embargo, en el vector alfabético, el elemento n-simo contiene un número que corresponde a la palabra n-sima en el orden alfabético. Este vector alfabético sirve de apuntador para secuencialmente emparejar los elementos lingüísticos del texto con el diccionario *WCL*.

Por ejemplo:

I/O AREA

Dirección:
0 2 8 11 21 31 34
Texto:
A pesar de recientes adelantos en la
Dirección:
37 48 51 54 67
Texto:
Tecnología de la documentación,

TABLA CORRESPONDIENTE AL I/O O AREA

Long.	Direc.	Semán.	Sintác.	Vector alfab.
1	0	B		0
5	2	J		4
2	8	J		2
9	11	A		8
9	21			10
2	31		P	5
2	34		A	6
10	37			9
2	48		P	1
2	51		A	3
13	54			7
1	67			11

Las frases que constituyen el abstracto se extraen del I/O AREA con un módulo llamado *semántico*, el cual incorpora más de un centenar de reglas que aplica jerárquicamente trabajando exclusivamente con las estructuras internosemánticas y los atributos sintácticos tal como aparecen en la TABLA.

El lector puede seguir en el apéndice la lógica del programa con un ejemplo práctico como el hacer un abstracto de las doce primeras sentencias de este mismo artículo.

APENDICE

PRODUCCION AUTOMATICA DE ABSTRACTOS

A pesar de recientes adelantos en la tecnología de la documentación, la producción de abstractos de alta calidad todavía tiene que hacerse manualmente. Por consiguiente, la preparación de abstractos y sus índices vienen a ser más de la mitad de los gastos en tiempo y en dinero por parte de las editoriales de revistas de abstractos. Así, pues, uno tiene que tener en cuenta factores de coste y tiempo cuando quiera comparar la producción de abstractos por medios manuales y automáticos.

Evidentemente existen dos alternativas como posibles soluciones al problema de la producción manual de abstractos:

1. Disponer de abstractos preparados por los propios autores como prerrequisito a la publicación de sus artículos.
2. Mecanizar la producción de abstractos.

En efecto, los editores de las revistas han tratado estos últimos años de convencer a los autores de incluir los abstractos. Sin embargo, el éxito ha sido muy variable, además de que estos abstractos han demostrado ser de un valor cuestionable. Por otro lado, la escasez de abstractores profesionales y factores de tiempo, costo y calidad han dado impulso hacia la investigación de métodos para producir abstractos e índices automáticamente. Así, pues, vamos a describir una técnica nueva para producir automáticamente abstractos de lo llamado «full text», es decir, a partir de documentos originales enteros.

Primero repasemos brevemente las funciones de un abstracto, que son las siguientes:

1. *Anunciamiento.*—Para dar alerta sobre la existencia de un documento.

TITULO: PRODUCCION AUTOMATICA DE ABSTRACTOS

Textos:

1. A pesar de recientes adelantos en la tecnología de la documentación, la producción de abstractos de alta calidad todavía tiene que hacerse manualmente.
2. Por consiguiente, la preparación de abstractos y sus índices vienen a ser más de la mitad de los gastos en tiempo y en dinero por parte de las editoriales de revistas de abstractos.
3. Así pues, uno tiene que tener en cuenta factores de coste y tiempo cuando se quiera comparar la producción de abstractos por medios manuales y automáticos.
4. Evidentemente existen dos alternativas como posibles soluciones al problema de la producción manual de abstractos.
5. 1. Disponer de abstractos preparados por los propios autores como prerrequisito a la publicación de sus artículos.
6. 2. Mecanizar la producción de abstractos.
7. En efecto, los editores de las revistas han tratado estos últimos años de convencer a los autores de incluir los abstractos.
8. El éxito, sin embargo, ha sido muy variable, además de que estos abstractos han demostrado ser de un valor cuestionable.
9. Por otro lado, la escasez de abstractos profesionales y factores de tiempo, costo y calidad han dado impulso hacia la investigación de métodos para producir abstractos e índices automáticamente.
10. Así, pues, vamos a describir una técnica nueva para producir automáticamente abstractos de lo llamado «full text», es decir a partir de documentos originales enteros.
11. Primero repasemos brevemente las funciones de un abstracto que son las siguientes.
12. 1. *Anunciamiento.*—Para dar alerta sobre la existencia de un documento.

TABLA 1

Sentencia	Palabra guía de mayor influencia	Valor semántico	Giro dado a la frase	Implicación	Resultado inicial sobre frase	Resultado final sobre frase
1	A pesar de	B	Comentario	Negativa	Desaparece	Desaparece
	Recientes	A	Histórico	Negativa	Desaparece	Desaparece
2	Por consiguiente	B	Explicación	Negativa	Desaparece	Desaparece
	Vienen a ser	A	Indefinido	Negativa	Desaparece	Desaparece
3	Así pues	B	Interpretación	Negativa	Desaparece	Desaparece
	Uno	A	Indefinido	Negativa	Desaparece	Desaparece
	Tener en cuenta	B	Conjetura	Negativa	Desaparece	Desaparece
4	Evidentemente	A	Opinión	Neg.	Desaparece	Desap.
	Posibles	B	Conjetura	Negativa	Desaparece	Desaparece
5	1.	Anáfora	? (Opinión) ←	? (Neg.) ←	?	? (Desap.) ←
6	2.	Anáfora	? (Opinión) ←	? (Neg.) ←	?	? (Desap.) ←
7	Ultimos años	A	Histórico	Negativa	Desaparece	Desaparece
8	Sin embargo	B	Comentario	Negativa	Desaparece	Desaparece
	Muy	E	Indefinido	Negativa	Desaparece	Desaparece
9	Por otro lado	B	Comparación	Negativa	Desaparece	Se queda ←
	Abstractos	G ←	(Título)	~ Positiva	Desaparece	Se queda
10	Vamos describir	I	Objetivo	Positiva	Se queda	Se queda
	Abstractos	G ←	(Título)	2 positiva	Se queda	Se queda –
	Espedir	H	Modificador	Nulo	Desaparece	Desaparece
11	Repasemos	A	Histórico	Neg.	Desap.	Desap.
	Brevemente	B	Limitador	Negativa	Desaparece	Desaparece
12	1.	Anáfora	? (Histórico) ←	? (Neg.) ←	? (Desap.) ←	? (Desap.) ←

La escasez de abstractos profesionales y factores de tiempo, costo y calidad han dado impulso hacia la investigación de métodos para producir abstractos e índices automáticamente.

Vamos a describir una técnica nueva para producir automáticamente abstractos de lo llamado «full text».

Nota.—Las sentencias 1-4 desaparecen por su peso negativo.

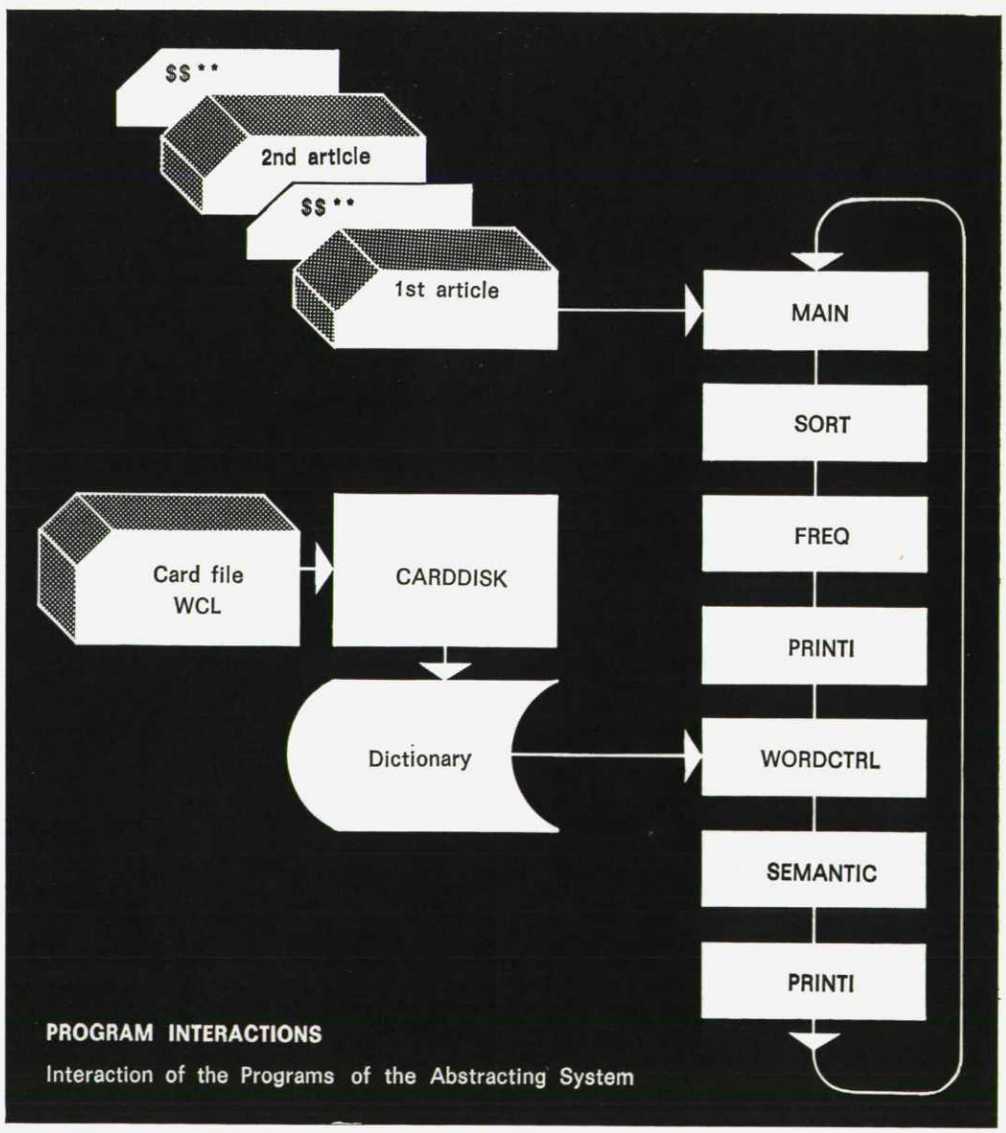
Las sentencias 5-6, por sí mismas nulas, se vuelven negativas por las anáforas «1» y «2», que las relacionan con la anterior, la 4.

Las sentencias 7 y 8 desaparecen por su peso negativo.

La sentencia 9, por sí misma suavemente negativa, tendría que desaparecer, pero está tan fuertemente influenciada por la 10 que se queda. Sin embargo el «por otro lado» de esta sentencia se elimina por ser expresión parentética.

La sentencia 10 indudablemente se queda por ser fuertemente positiva, pero el «Así pues» y «, es decir a partir de...», también se eliminan por ser expresiones parentéticas.

La sentencia 11 se va por negativa, y la 12, ella misma nula, también desaparece por el anáfora «1».



PROGRAM INTERACTIONS

Interaction of the Programs of the Abstracting System