AUTOMATIZACIÓN DE LOS DICCIONARIOS DE SINÓNIMOS

Dentro del campo de las aplicaciones de la Informática a la Lingüística existe una serie de problemas que ya han sido, básicamente, resueltos: automatización general, lematización, frecuencias y rango, estilométrica, etc. (No sólo en lo que afecta a los idiomas extranjeros —inglés, francés..., sino también en lo que afecta a la lengua castellana: hace cinco años, más de quince Universidades extranjeras habían estudiado los mencionados problemas, y habían indexado textos castellanos de todas épocas y de todo género [USA, Canadá, México, Francia...]).

Dentro de los puntos esenciales que siguen problemáticos, citaríamos la sintaxis automatizada —que fue objeto de una ponencia nuestra hace cuatro años en Barcelona [1]— y la traducción automática, que dejamos de lado, para fijarnos más precisamente en uno de los aspectos de la semántica.

Hay que recordar, aquí, entre otros intentos que se acercan a nuestro propósito, el llevado a cabo por los colaboradores del equipo del T. L. F. *(Trésor de la Langue Française)* en Nancy (Francia), que han estudiado los grupos constituidos por una misma categoría gramatical, unidos por un conector común. (Por ahora se han trabajado los grupos —*Adjetivo* «OU» *Adj.*—, aunque se prevé la implicación de otros conectores. No hay que subrayar cómo, de este tipo de estudio, se pueden sacar conclusiones que interesan a la sintaxis pero también muy especialmente a la semántica. Sobre estas experiencias cabe pro-

[1] E. Moreu-Rey, M. A. Vidal, «Posibilidades de aplicación mecánica a las estructuras sintácticas», ponencia del Congreso «Informática y Lingüística», Madrid (Fundesco), 1977.

ceder a la aplicación de la teoría de los grafos: en Nancy lo han experimentado y a las publicaciones correspondientes me remito [2].

Las unidades agrupadas por fr. *ou* (cast. *o*); o también por fr. *et* (cast. *y*) —conjunción prevista igualmente en Nancy, presentan una equivalencia, una oposición, una gradación... Ahora bien, puede ocurrir que las dos unidades ligadas por este conector posean simplemente un sema común, no quedando, por tanto, la relación semántica indicada con la indispensable precisión.

Por nuestra parte hemos intentado utilizar otros conectores que, aunque aparecen con menor frecuencia en el discurso que los estudiados en Nancy, ofrecen la ventaja de definir mejor el grupo binario: una mayor seguridad en la interpretación de la asociación u oposición de las unidades confrontadas.

En mis trabajos, mi punto de partida ha sido la locución adverbial *à la fois* — ET —, que corresponde más o menos al español *a la vez* — Y — ('a la par, 'al mismo tiempo'): 'tonto y malo a la vez'; 'a la vez tonto y malo'... por ejemplo.

El programa (realizado por mí en lengua PL/I en Barcelona, y bajo mi dirección, en COBOL-ANS, el que ha sido aplicado al corpus capitalizado en Nancy) permite obtener las secuencias de sustantivos, adjetivos, verbos, reunidos por este conector preciso, tomado aquí como ejemplo:

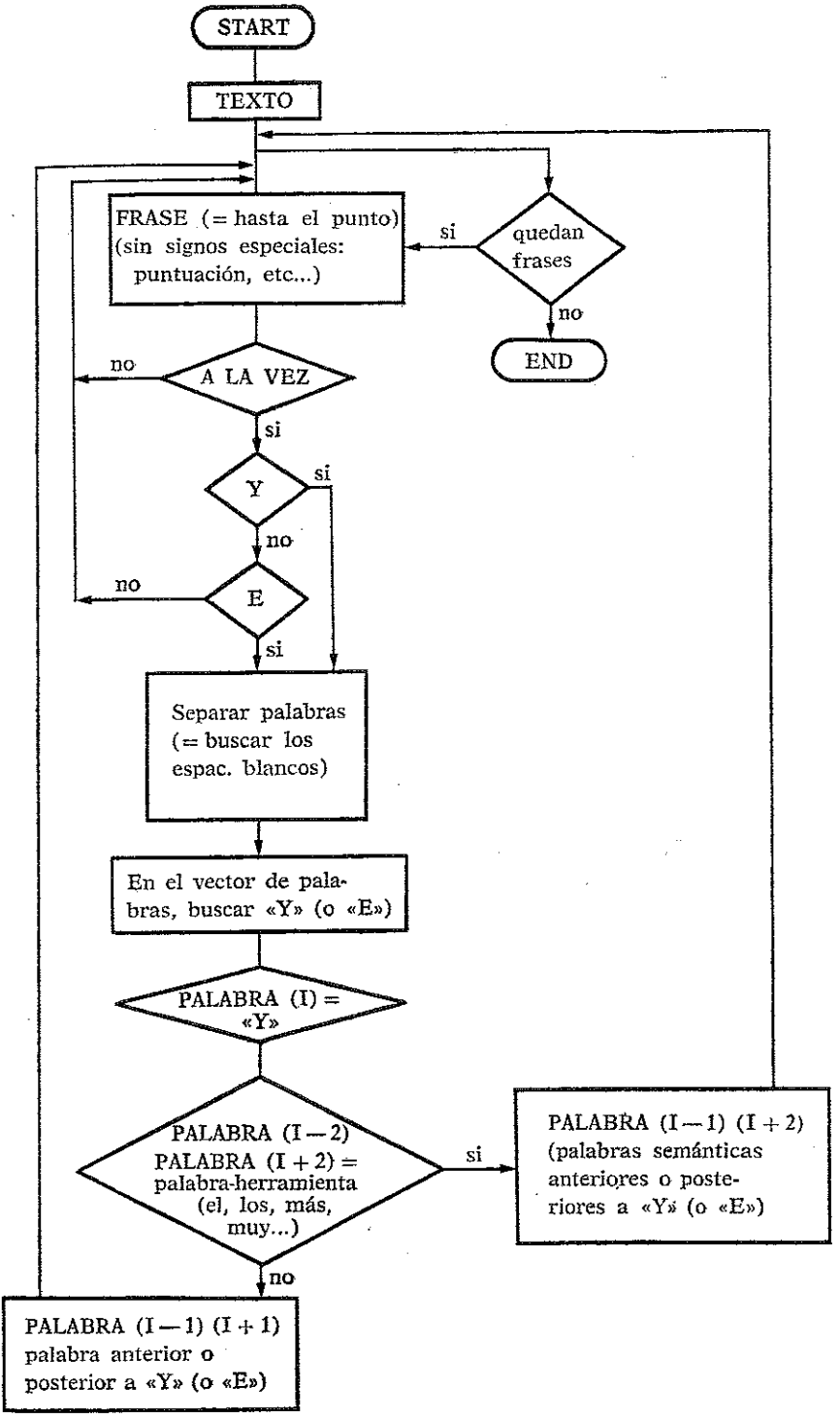
En síntesis el organigrama viene a ser el siguiente:

— La máquina procederá a la lectura del texto; leerá la primera frase. (Se considera la frase como unidad para analizar, en este caso.) La máquina buscará entonces el punto. (Dejamos de lado problemas tangentes como el de la puntuación.)

— El ordenador buscará la construcción *a la vez,*
  — si no aparece, el ordenador pasa a la frase siguiente;
  — si aparece, procede entonces a la localización de la conjunción *y* (o *e* —en castellano, donde el programa se presenta ligeramente más complicado— esta operación suplementaria no siendo necesaria para el francés: *casado e infeliz a la vez).*

---

[2] Michon-Potdevin, «Théorie des graphes», *Le Français Moderne* 4, Oct. 1973, Nancy. y J. Y. Hamon, J. P. Michon, «La théorie des graphes en linguistique mathématique», *Le Fr. Mod.* (supplément, fasc. 1), 1974.


START
TEXTO
FRASE (= hasta el punto) (sin signos especiales: puntuación, etc...)
si
quedan frases
no
END
no
A LA VEZ
si
Y
si
no
no
E
si
Separar palabras (= buscar los espac. blancos)
En el vector de palabras, buscar «Y» (o «E»)
PALABRA (I) = «Y»
PALABRA (I — 2) PALABRA (I + 2) = palabra-herramienta (el, los, más, muy...)
si
PALABRA (I — 1) (I + 2) (palabras semánticas anteriores o posteriores a «Y» (o «E»)
no
PALABRA (I — 1) (I + 1) palabra anterior o posterior a «Y» (o «E»)

— si no hay conjunción, ello supone que no existe grupo binario y que se trata, por ejemplo, de una frase del tipo *hablaban todos a la vez:* frases que no retiene la máquina.

— Si la construcción *a la vez* y la conjunción *y* están presentes en la frase estudiada, se constituye un vector de palabras (entendiendo por *palabras* lo encuadrado entre dos espacios, entre dos blancos).

— Luego la máquina retendrá e imprimirá el vocablo anterior y el vocablo posterior a la conjunción ($y + 1$ / $y - 1$): la frase queda igualmente retenida.

(Se supone el ordenador provisto de un «diccionario» para la eliminación de los morfemas o palabras-herramienta (artículos, preposiciones...) con el fin de considerar, únicamente, las formas-«llenas» (sustantivos, adjetivos, verbos.)

En la confección del programa se han presentado, evidentemente, algunas dificultades, que no cabe enumerar ahora (frases complejas con doble conjunción; y otros tipos que se dan, en realidad, en raras ocasiones). Pero el programa resuelve la gran mayoría de los casos —y estamos estudiando las mejoras que puedan resolver más del 90 % de los ejemplos reales.

La rentabilidad de la mencionada estructura con —*à la fois ... et*— viene proporcionada por los datos siguientes: —aplicado el programa a una cuarta parte del corpus total listado en Nancy (que comprende textos literarios de los siglos XIX y XX —es decir 20 millones de ocurrencias sobre un total de 80 millones) nos ha permitido la obtención, ya, de una suma de 625 grupos binarios de adjetivos (sólo de adjetivos) que proporcionan 735 adjetivos diferentes. Resultado no despreciable teniendo en cuenta que *a la vez* —*y*— o *à la fois* —*et*— no pertenecen precisamente al grupo de las estructuras más frecuentes.

El programa, en este caso concreto el análisis del conector *a la vez* —*y*—, recoge, pues, una lista de grupos binarios (de adjetivos, sustantivos, verbos) en oposición sintagmática, marcando: ya sea una oposición (caso que se presenta con mayor frecuencia —los dos términos son antónimos); ya sea una equivalencia o una gradación... Es decir que, aplicado este programa a corpus muy extensos (del orden de más de cien millones de ocurrencias si queremos que empiece a cobrar una indiscutible rentabilidad) establecería una lista de

antónimos (principalmente), sinónimos, o hipónimos, hallados realmente en los enunciados.

A partir de estas experiencias realizadas ya, nos es lícito adentrarnos en una segunda parte prospectiva.

Este complemento a la primera etapa consistirá en pedir a los ordenadores el registro de los resultados obtenidos con programas similares al precedente, con el propósito de obtener un inventario exhaustivo de Grupos Binarios más o menos «orientados».

Para ello, hemos estudiado toda una serie de estructuras o moldes sintáctico-lógicos (conectores) muy diversos: preposiciones, conjunciones, locuciones, verbos, sustantivos, o la puntuación misma, que no deja de ofrecer dificultades, de las que se ha hecho mención.

Teniendo en cuenta el funcionamiento de cada uno de los conectores se preverán programas con un triple efecto —y cuyos fundamentos se basarán en el cálculo de probabilidades, resuelto partiendo de combinatorias.

Se considerarán en primer lugar aquellos conectores que proporcionen listas de Grupos Binarios comprendiendo sin discriminación antónimos y sinónimos. Ejemplos:

a O b; a Y b; SEA a SEA b; NI a NI b; MAS a QUE b; etc.

En segundo lugar, los que no ofrecerían más que sinónimos o casi sinónimos:

a ES DECIR b; a A SABER b; a NO ES MAS QUE b («la felicidad no es más que el placer compartido»).

Citamos solamente algunos de los ejemplos previstos.

Y, en tercer lugar, aquellos que seleccionarán normalmente antónimos:

a AL CONTRARIO b; a POR EL CONTRARIO b; NO a SINO b; ALTERNANCIA DE a Y DE b; DISOCIACIÓN DE a Y b; MEZCLA DE a Y b; a NO ES b; a AUNQUE b; a PERO TAMBIÉN b («conocí el placer pero también la angustia»...).

Las muestras precedentes bastan para dar una idea clara del propósito [3].

[3] El estudio de los conectores dio pie a capítulos enteros de nuestra tesis doctoral: *Méthodologie pour l'étude des champs sémantiques. Une application: le réseau lexico-conceptuel de plaisir.* Setiembre 1979. Barcelona, Facultad de Filología. Ms. multicopiado (Resumen impreso).

Con herramientas de entre las que preceden, o similares, obtendríamos así secuencias mayoritarias del tipo:

1. *feliz — dichoso,* en el grupo de los programas que operan con los conectores que proporcionan principalmente sinónimos o casi-sinónimos, y que alcanzarían un 60 % de promedio (55 % en una serie, 70 % en otra...).

2. *feliz — infeliz; feliz — desgraciado* —con un 60 %, por medio de los moldes que seleccionan antónimos (p. ej. NO *a* SINO *b*...).

No dejan de aparecer, al margen del grupo precedente, resultados superficialmente paradójicos:

dichoso — contento
desgraciado — descontento
descontento — inquieto,

hallados en relación antagónica (del orden del 10 % al 15 %). P. ej. «No dichoso sino contento».

Para clarificar ciertos resultados que se dan con una frecuencia intermedia se recurrirá a conectores que indican la gradación:

> *NADA SE PARECE A a COMO b; a SE PARECE A b; a NO ESTA LEJOS DE b; PREFERIR a A b; a SIN LLEGAR A b; a NO LLEGANDO A b; NO TAN a COMO b; MAS a QUE b; a CASI DE b* («momentos de alegría, casi de felicidad»...).

En este mismo sentido, algunos diccionarios empiezan ya a recoger y a indicar los sinónimos en gradación (hipónimos, hiperónimos...)[4].

3. Para terminar, no hay que desechar una cuarta serie de resultados: (aunque cuantitativamente menores aún: del orden del 5 % a 7 %).

alegría — cielo
alegría — jardín
alegría — azul
felicidad — amor
dulzura — seda

---

[4] E. Genouvrier, C. Désirat, T. Hordé, *Nouveau Dictionnaire de Synonymes,* París, Larousse, 1977. (Por ejemplo: «affection ↑ amour ↓ amitié».) Pretendemos llegar al estadio donde estos resultados puedan ser obtenidos mecánicamente.

Esta serie serviría, al menos, para introducir en los diccionarios (que los ignoran por lo general) las connotaciones más generalmente adoptadas en el habla, y que han alcanzado el rango de verdaderas denotaciones.

Todos los ejemplos que preceden quedarán automáticamente localizados y procesados. No así el conjunto de casos «extravagantes» que serían rechazados por el ordenador cuando no alcanzaren frecuencias mínimas.

En suma, un procedimiento manual serviría ciertamente para separar sinónimos de antónimos, pero pretendemos que tal resultado puede ser alcanzado por procedimientos mecánicos, según los porcentajes del tipo de relación (equivalencia, oposición, gradación...) que ofrecen cada uno de los conectores previstos.

Sin contar que los trabajos que estamos llevando a cabo pretenden —entre otros objetivos: la realización y la mejora objetiva de los diccionarios analógicos e ideológicos, y por ende de los diccionarios generales de la lengua, así como de los diccionarios bilingües o de traducción, los cuales —grosso modo y con correcciones mínimas debido al heteromorfismo— deberían ser el resultado automático de la confrontación de los diccionarios analógicos de cada una de las lenguas examinadas.

La Informática, en estos últimos años, viene realizando progresos extraordinariamente rápidos. Ello nos permite imaginar un recurso casi inmediato a los ordenadores con lectores ópticos, que proporcionarán resultados en cantidades diez o más veces superiores a las actuales: por ejemplo de los 500 a 800 millones de ocurrencias (teniendo en cuenta lo recogido en Nancy con los ordenadores de la antigua generación).

Estos progresos ofrecen perspectivas que posibilitarán la realización práctica de los programas aquí previstos —meta que, desde ahora, ya no puede considerarse utópica.

MARÍA ANGELES VIDAL COLELL